UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

University of Padova

DEPARTMENT OF INFORMATION ENGINEERING

MASTER'S DEGREE IN ICT FOR INTERNET AND MULTIMEDIA

Joint Polarization and Event Sensing for Surface Normal Reconstruction

Supervisor:
Prof. Pietro Zanuttigh
(University of Padova)

Student:
Anxhelo Kamberi

Co-Supervisor :
Alexander Gatto
(Sony Europe B.V.)

ACADEMIC YEAR: 2021-2022
Date of graduation: 11/07/2022

Abstract

Surface normal reconstruction from combining linear polarization and standard monochrome cameras has been widely studied in the past years. The aim of this work is to study for the first time the possibility of reconstructing the surface normals of a scene from the data collected by exploiting a linear polarizing filter and an event sensing camera, where the latter is a novel sensor whose applications are still an open research in the computer vision field. For this task a deep learning algorithm has been used in order to perform the conversion of the event camera's signal to a representation resembling the one acquired by a standard monochrome camera, then the surface normals are reconstructed with an already existing procedure.

Abstract

La combinazione di fotocamere monocromatiche e di filtri linearmente polarizzati per la ricostruzione delle normali delle superifici è stata ampiamente studiata negli ultimi anni. Lo scopo di questo lavoro è di provare a sperimentare per la prima volta la possibilità di ricostruire le normali delle superfici in una scena a partire da dati raccolti tramite un filtro linearmente polarizzato e un nuovo tipo di sensore noto come *event-sensing camera*, le cui applicazioni nel mondo della visione computazionale sono ancora un campo di ricerca aperto. Per questo scopo, è stato implementato un algoritmo di intelligenza artificiale al fine di eseguire la conversione del segnale acquisito con la event-sensing camera in una rappresentazione simile al segnale acquisito da una fotocamera monocromatica tradizionale, in modo da ricostruire successivamente le normali delle superifici attraverso una procedura già esistente.

Contents

Acronyms

- *TRS* : Transmitted Radiance Sinusoid
- *DOP* : Degree of Polarization
- *RGB* : Red, Green, Blue
- *DVS* : Dynamic Vision Sensor
- *APS* : Active Pixel Sensor
- *DAVIS* : Dynamic and Active pixel VIsion Sensor
- *AER* : Address-Event Representation
- *jAER* : JavaAddress-Event Representation

1 Introduction

Polarization has proven to be a useful source of information in the analysis of light scattering from surfaces in computer vision. There are a number of ways in which polarization arises and a considerable amount of literature has investigated the use of polarization for surface analysis. Most of the research aimed at extracting and interpreting information from polarization data by placing a linear polarization filter in front of a traditional monochrome camera and taking images of an object or a scene with the polarization filter oriented at different angles, in order to produce a set of polarization images to be analyzed [1]. However, also liquid crystal technology [2] has been exploited in order to acquire multiple polarization images per second, and polarization filter array cameras [3] have been used for getting in a snapshot a set of polarization images orientation in order to speed up the polarization images acquisition.
This work aims at investigating for the first time the possibility of reconstructing the surface normals of a scene by using a standard setup with a polarization filter rotated in front an event-sensing camera rather than a traditional monochrome camera. Event sensing cameras are novel sensors inspired by the biological functioning of the human retina that encode only changes in light intensity instead of acquiring absolute intensity values as in traditional cameras, and have independent pixels for avoiding redundant data, as presented in Chapter 3. The main advantages of the event cameras over the standard ones are higher dynamic range, low power consumption and higher acquisition rate. For this reasons it seems interesting to make a preliminary exploration on the behaviour of the event sensing technology for surface analysis from polarization cues in order to benefit from the advantages of the event cameras for possible real time applications, especially in the robotics field.
The pipeline for this work is shown in Figure 2, and it consists in two main steps. The first one, which is the most important, consists in acquiring polarization information with an event sensing camera while a polarization filter is rotated in front of it, then develop a deep learning algorithm in order to pixel-wise transform the data collected from the event camera into a representation resembling the ones acquired by a standard monochrome camera. In the last step, the surface normals are reconstructed according to a standard literature method [4] and by using the data outputted by a deep learning algorithm rather than the output of a traditional camera.

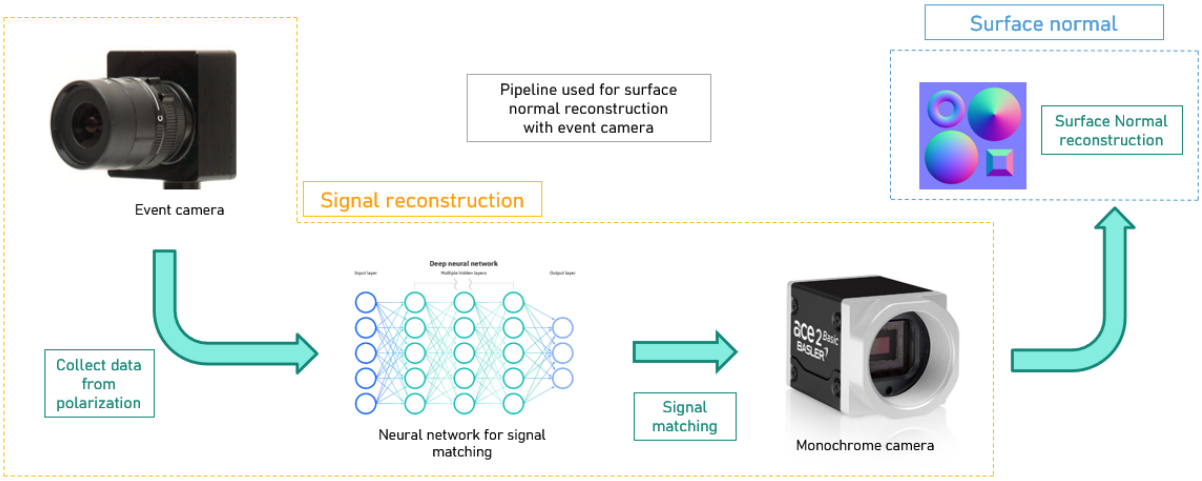


Figure 2: Graphical illustration of the pipeline used throughout this work.

2 Surface Normal Reconstruction from Polarization Imaging

2.1 Light Polarization

Polarization is a property applying to transverse waves such as light that specifies the geometrical orientation of the waves oscillations. So, polarization of light refers to the direction of the electric field oscillation of the light. The electric field can be decomposed in two main directions of oscillation, and so three different types of light polarization can be defined: linear polarization, where the fields oscillate in a single direction as the wave travels, circular polarization, where the fields oscillates at a constant rate along the two main directions as the wave travels, and elliptical polarization, which is a more general case of circular polarization. For this work, linear polarization of the light is exploited for the purpose of reconstructing the surface normals orientations for dielectric objects of a given scene.
Light can be linearly polarized by using a linear polarizing filter, which makes the randomly oscillating electric field directions of unpolarized light to oscillate only along a specific direction, as illustrated in Figure 3.

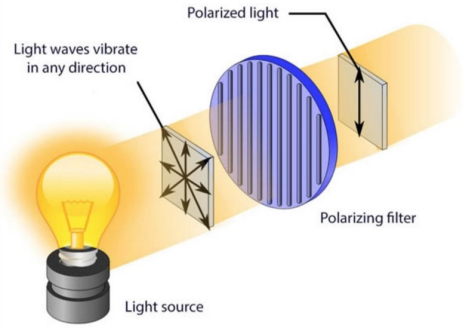


Figure 3: Illustration of the working principle of a linear polarizing filter.

Another method for polarizing light is by specular reflection: the light reflected by an object is partially polarized, i.e. consists of an unpolarized component and a completely polarized component, along the oscillating direction parallel to the surface of the object, as illustrated in Figure 4. From this fact, it is understandable that there is a relationship between the normal vector of a surface, defined as the vector which is perpendicular to the surface at a given point, and the partially polarization state of the light reflected by a surface.

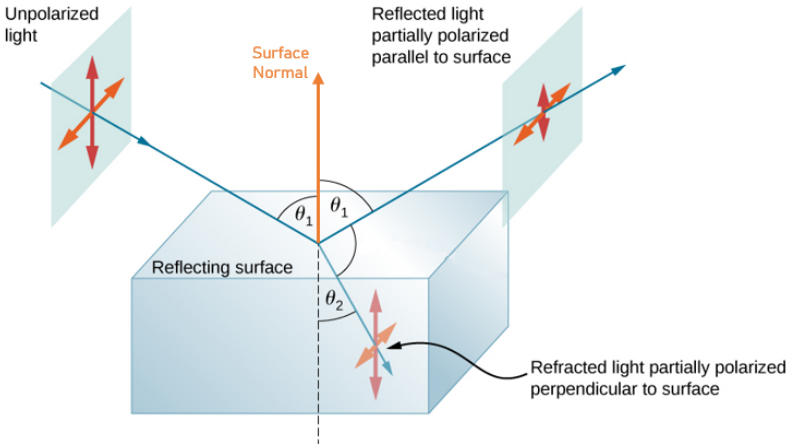
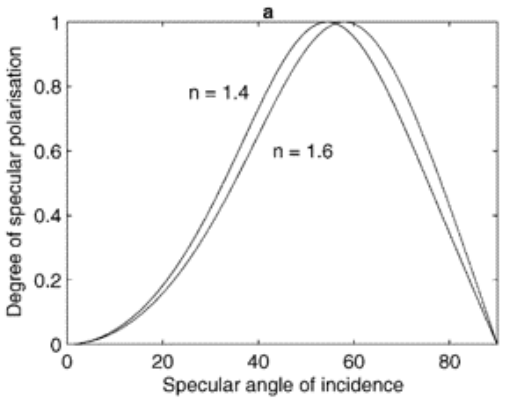
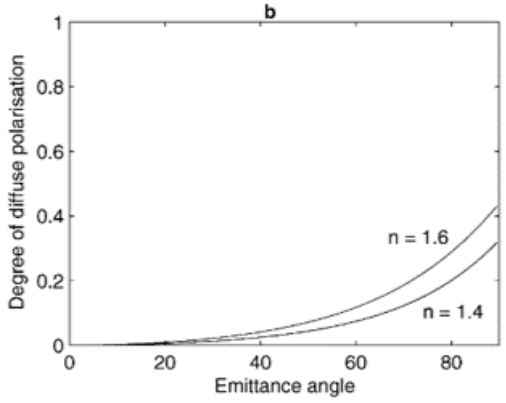


Figure 4: Illustration of the working principle of partial polarization of light by specular reflection.

The quantity used to describe how much the light is polarized is named *Degree Of Polarization* (DOP), and it assumes a value between 0 and 1. The higher the DOP, the more the light is partially polarized. For dielectric materials, the DOP of specular reflection in function of the incidence angle of the light is shown in Figure 5a, where the incidence angle of the light is defined as the angle between the light direction and the surface normal vector, and it reaches it maximum value at the socalled *Brewster's angle* of approximately $60°$.



(a) Degree of polarization of specular reflection at the variation of the light incidence angle for dielectric materials.



(b) Degree of polarization of diffuse reflection at the variation of the light emittance angle for dielectric materials.

Figure 5: Degree of polarization for dielectric materials for different values of refractive index n.

Furthermore, the DOP of reflected light depends also on the *refractive index* of the object reflecting the light, but since for dielectric materials the refractive index n belongs to the range $n \in [1.4, 1.6]$ it is possible to approximate the refractive index of dielectric materials with the value $n = 1.5$. For smooth dielectric surfaces, also the diffuse reflection is polarized as well [4]. The light hitting a surface is partially absorbed, then it undergoes into a scattering process and part of the scattered light is then re-transmitted back to the initial medium as diffuse reflection, as Figure 6 illustrates. For diffuse reflections, the DOP depends on the emittance angle of the diffused light, defined as the angle between the surface normal and the emitted light direction. The key difference between the diffuse and specular reflection is the DOP: diffuse reflection has on average a lower DOP compared to specular ones, as by comparing Figures 5a and 5b.

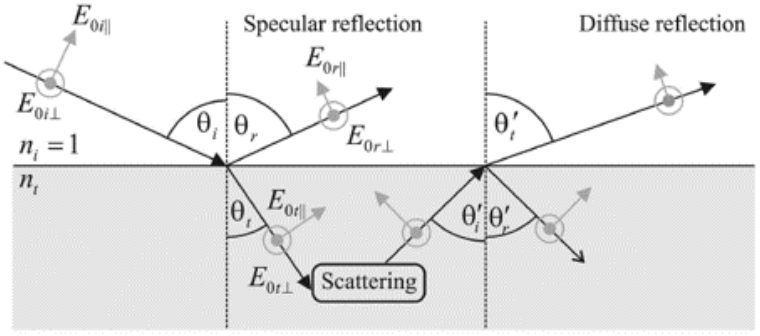


Figure 6: Illustration of the working principle of partial polarization of light by diffuse reflection.

2.2 Surface Normal Reconstruction methods

For the surface normal reconstruction task from polarization cues of dielectric materials, several approaches already exists [5], [6], [7], [8], [2], but they require a more complex setup compared to the one used for this work. For this reason, in this work the surface normal reconstruction task is performed by following the method proposed in [4] that is described in this section.

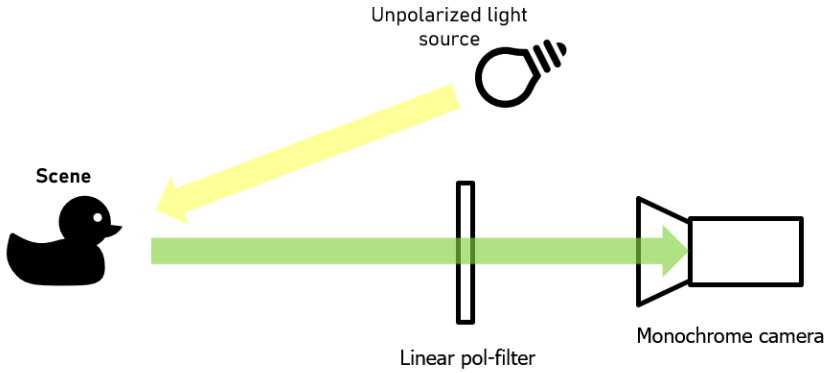


Figure 7: Schematic illustration of the setup used for surface normal reconstruction in [4].

By considering a setup with a linear polarizing filter placed between a scene and a monochrome camera as in Figure 7 and the theory of polarization by reflection described in 2.1, as the polarization filter is rotated, the measured brightness at a given pixel of the monochrome camera varies according to the socalled *transmitted radiance sinusoid* (TRS):

$$I(\theta_{pol}) = \frac{I_{max} + I_{min}}{2} + \frac{I_{max} - I_{min}}{2}\ cos(2\theta_{pol} - 2\phi) \tag{1}$$

where I_{max} and I_{min} denotes are the maximum and minimum observed pixel brightness values as the filter is rotated, θ_{pol} is the angle which the rotated polarization filter makes with the initial vertically upwards orientation, ϕ and is the phase angle of the sinusoid. The TRS for a given pixel can be computed by only capturing a set of polarization images at $0°$, $45°$, $90°$, $135°$ polarization filter angles, as shown in Figure 8.

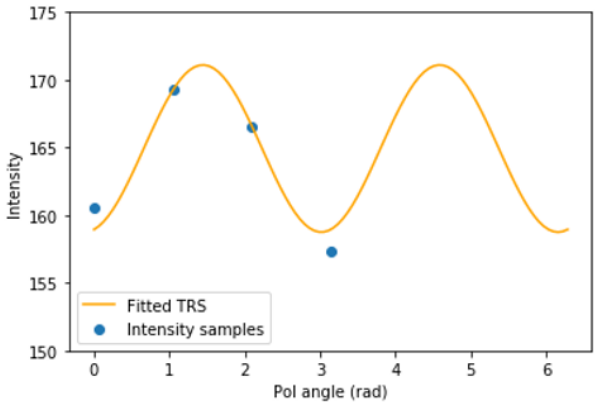


Figure 8: Example of TRS for a given monochrome camera pixel, computed with the pixel intensities at the polarization filter angle of $0°$, $45°$, $90°$, $135°$.

By considering the *Azimuth* angle of the surface normal as the angle of the projection of the normal onto the image plane relative to a reference coordinate system and the *Zenith* angle of the surface normal as the the angle between the surface normal and the viewing direction of the camera, it is possible to fully characterize the surface normal orientation by computing these two angles. Since the DOP can be also computed from the TRS information as:

$$DOP = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \tag{2}$$

with the method proposed in [4], the Zenith angle of the surface normal associated to a given pixel can be directly computed from the DOP: for diffuse reflection pixels by inverting the DOP function of Figure 5b, while for specular reflection pixels by inverting the DOP function of Figure 5a. For specular reflection pixels, from Figure 5a it is possible to see that for a given DOP value, there exists two Zenith angle solutions, except at the Brewster angle, leading to problem known as *Zenith angle ambiguity* which can be overcome by using a setup in which diffuse light is reflected by the surfaces of a given scene.

The Azimuth angle of the surface normal instead corresponds for [4] to the polarization filter angle at which I_{max} is reached, in case of diffuse reflection, and to the polarization filter angle at which I_{min} is reached, in case of specular reflection. Thus, for both specular and diffuse reflection the Azimuth angle has two possible solutions since the TRS has two maxima and two minima in a full 360° rotation, due to the characterization of the linear polarizing filters. Since the Azimuith angle denotes the angle of the projection of the normal onto the image plane relative to a reference coordinate system, the Azimuth angle ambiguity is also known as *convex/concave ambiguity*, for the reason illustrated in Figure 9.

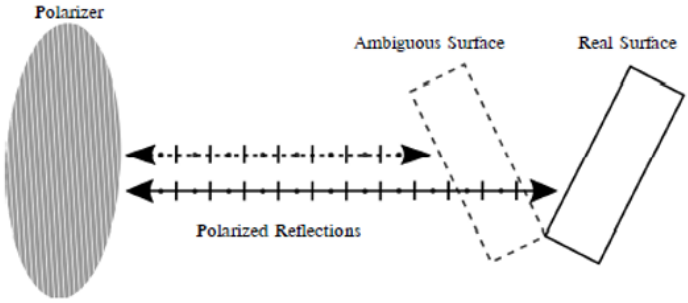


Figure 9: Illustration of convex/concave ambiguity of a surface due to the Azimuth angle ambiguity.

The Azimuth angle ambiguity are solved in [5], [6], [7], [8], by using different methods such as acquisition of polarization images from multiple views and acquisitions with multiple light source positions. However, these methods are beyond the scope of this work.

By denoting the Azimuth angle as α and the Zenith angle as β, the surface normal orientation can be computed in Cartesian coordinates as:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \sin(\beta)\cos(\alpha) \\ \sin(\beta)\sin(\alpha) \\ \cos(\beta) \end{pmatrix}$$

which can be encoded then in a RGB representation for visualization purpose.

An block illustration of the pipeline used in this work for reconstructing the surface normal orientation is shown in Figure 10.

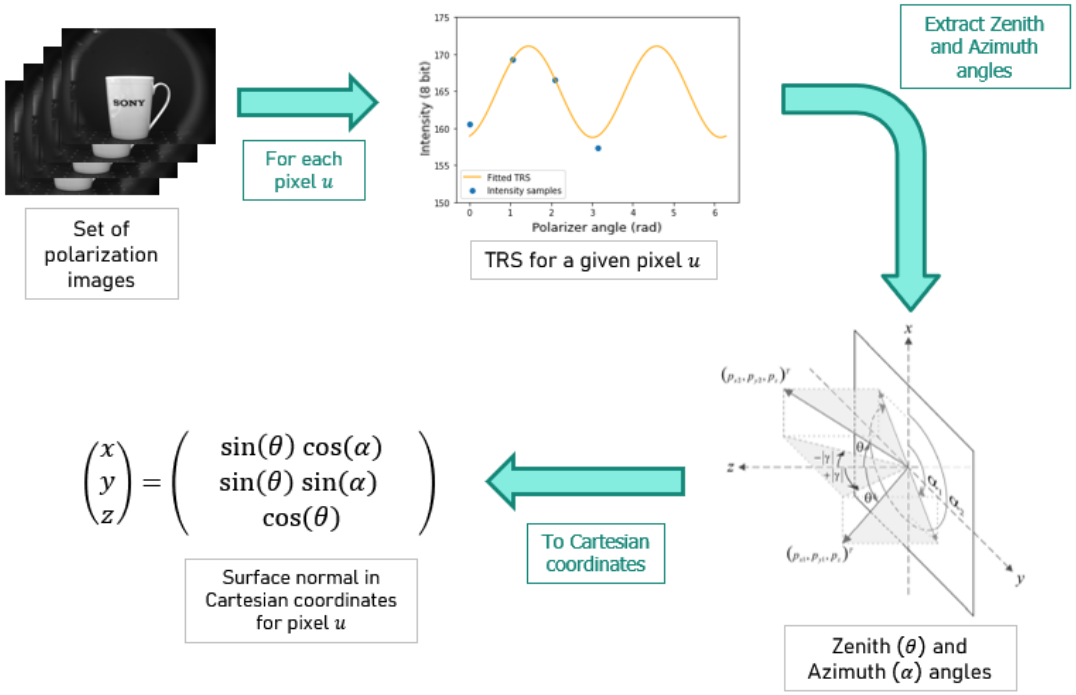


Figure 10: Block illustration of the pipeline used for surface normal reconstruction.

3 Event Sensing

3.1 Event sensing cameras

Event sensing cameras are bio-inspired sensors that pose a paradigm shift in the way visual information is acquired, offering numerous advantages when compared to standard camera systems [9], [10] but also some drawbacks. Traditional frame-based image sensors capture light information based on a fixed clock that has no relation to the dynamics of the viewed scene, giving so the same importance to static and dynamic parts of the frames. This working principle is not efficient, since it causes conventional cameras to produce a huge amount of redundant information in several frames, requiring so a good computational power and an efficient data processing for dealing with real time applications. To overcome this limitations and to abandon the concept of constant frame-rate vision, the branch of *neuromorphic vision engeneering* has aimed to built neuromorphic vision sensors which tries to mimic the biological working principle of the human retina [11] [12], in order to process the visual information in a more efficient way. The event-based cameras, which are an evolution the first neuromorphic sensors known as *silicon retinas*, relies on a smart and efficient sensors that create events rather than images by asynchronously measure the per-pixel brightness changes, and output a stream of events that encode the time, pixel coordinates and sign of the brightness change. Figure 11 briefly illustrates a comparison between the output of a traditional camera and the event-sensing one.

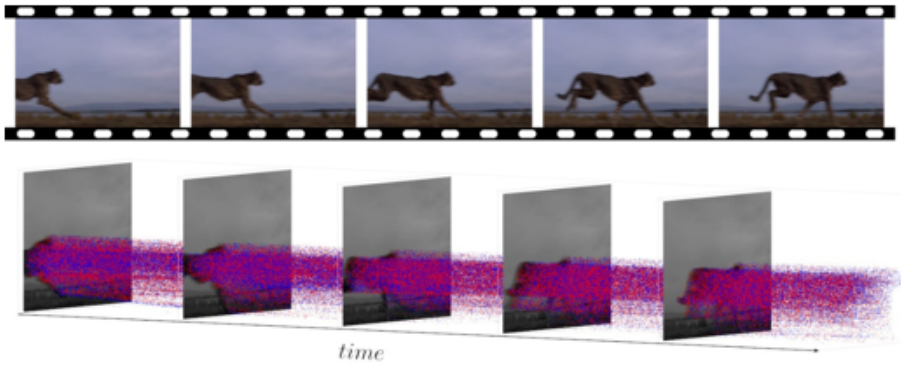


Figure 11: Output comparison between standard cameras and event sensing camera.

The key advantages of event sensing cameras, compared to conventional ones are: high temporal resolution (in the order of μs) and high pixel bandwidth (on the order of

kHz), resulting in a reduced motion blur, very high dynamic range ($140\,dB$ vs. $60\,dB$) and low power consumption, making the event sensing cameras suitable for robotics and real time applications.
Over the last two decades, three main types of event camera pixel designs have been developed: the *Dynamic Vision Sensor* (DVS) [13], *Asynchronous Time Based Image Sensor* (ATIS) [14] and *Dynamic and Active Pixel Vision Sensor* (DAVIS) [15] [16]. The DVS event sensor contains pixels that are only capable to measure only log-scale light intensity changes, without having the possibility to acquire an absolute brightness value as in traditional cameras.

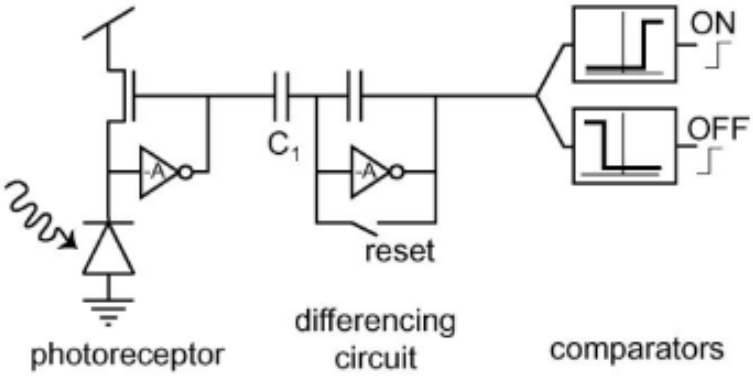
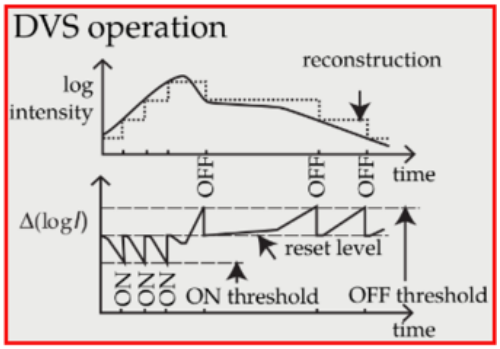
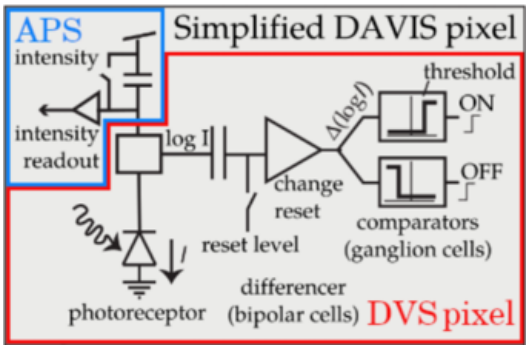


Figure 12: Simplified schematic of the DVS pixel circuitry.

In fact, as illustrated in the DVS pixel circuit abstraction of Figure 12, the continuous-time photoreceptor transforms the light intensity into a logarithmic-scale voltage V which is then compared to a reference voltage V_0: a socalled *ON* event is outputted if $V - V_0 \geq ON_T$, meaning that a light intensity increase is detected, or a socalled *OFF* event is outputted if $V - V_0 \leq OFF_T$, meaning that a light intensity decrease is detected. In the DVS sensor, the threshold values ON_T and OFF_T can be set as desired, which for the aim of this work is an essential feature for acquiring data related to polarization as it will be discussed in Chapter 4.2. The working principle of the DVS pixel is briefly illustrated in Figure 13a, where changes in log-scale intensity with respect to the corresponding thresholds implies the production of ON or OFF events. The widely used DAVIS event camera instead, as the ATIS, combines a conventional CMOS active pixel sensor (*APS*) in the same pixel with DVS as illustrated in Figure 13b, which allows for acquiring grayscale images as in standard monochrome cameras. The APS frames of the DAVIS sensor are exploited in this work for reconstructing the TRS from the events, as it will be discussed in subsection 6.2.2.



(a) Operation of a DVS pixel, converting light intensity changes into events.



(b) Simplified schematic circuitry of DAVIS pixel.

Figure 13

DVS Resolution	346 x 260 pixels
Frame Resolution	346 x 260 pixels, Grayscale, simultaneous output with DVS
DVS Dynamic range	120 dB
APS Dynamic range	56.7 dB
Min. latency	~ 20 us
Lens mount	CS-mount
Connectors / Power	USB 3.0 micro
Bandwidth	12 MEvents / second
Software	DV-Platform
Power consumption	< 180mA @ 5V DC
Dimensions	H 40 x W 60 x D 25 [mm]
Weight	100g (without lens)
Hardware multi-camera sync	Supported (HiRose Connector)
IMU	6-Axis Built-in
Case	Anodized aluminum, 4 mounting points
Tripod mount	Whitworth ¼" female
APS Frame Shutter	Configurable, Global or Rolling Shutter
CMOS Technology	0.18 um 1P6M MIM CIS
Chip size	8 x 6 [mm]
Pixel size	18.5 x 18.5 [um]
Array size	6.4 x 4.8 [mm]
Fill factor	22 %
Pixel complexity	48 transistors, 2 capacitors, 1 photodiode with micro-lens
Chip voltages	1.8 V and 3.3 V
Chip power consumption	DVS: 10-30mW (activity dependent) APS: 140mW
APS dark signal	18000 e^-/s
APS readout noise	55 e^-

Figure 14: DAVIS-346 RED specifications sheet.

Moreover, the event-sensing camera used during this work is the DAVIS-346 RED, which specifications sheet of the camera [17] is reported in Figure 18.

3.2 Events representation

The address-event representation (*AER*) is an asynchronous event communication protocol widely used in neuromorphic engineering and a key building block to event-based vision sensors, which allows the event sensors to communicate with external architectures. In the AER protocol each event E outputted by the event-sensor is encoded as a tuple:

$$E = (t, x, y, p)$$

where t denotes the timestamp in μs resolution at which the event E has occurred, useful information if the data is not processed in real time but at a later time, x and y represents the coordinates of the pixel that fired the event E, and $p \in \{-1, 1\}$ represents the socalled *polarity* of the event, which assumes value $p = 1$ in case of ON event and value $p = 0$ in case of OFF event.
The *Java-Address Event Representation* (jAER) [18], a Java-based framework exploiting the AER protocol for event-camera data transmission, allows for the visualization of the event-sensor output in real time. Furthermore, the jAER software allows for tuning the event-camera parameters, such as the ON/OFF events thresholds, the *refractory period*, which is the pixels response time between inter-events, and the sensor's bandwidth, in order to set the parameters of the event sensor for the desired application.

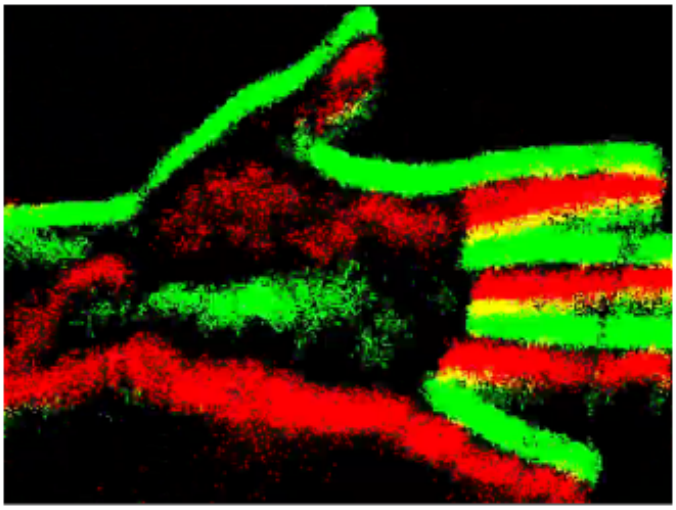


Figure 15: Screenshot of the jAER events real-time representation of the DVS events outputted at a fixed time. On the scene events from a moving hand are displayed: ON events are denoted by red pixels, OFF events by red ones.

This software is used during this work for both evaluating the behavior of the DAVIS-346 RED to changes in light intensity related to polarization at different parameter values of the camera, and also for acquiring data, as furthermore discussed in the 4.2 [18]. An example of the real-time event visualization with jAER software is shown in Figure 15, where ON events are denoted by green pixels, and OFF events are denoted by red pixels.

3.3 Challenges and Applications

The developement of the event cameras has lead to the challenge of designing novel methods to process the acquired data and extract information from it in order to unlock the advantages of the camera. The main challenges of this novel vision paradigm are related to coping with different space-time output, different photometric design, and with noise and dynamic effects of the event-cameras. In fact, as described in the previous sections, the output data of event cameras is asynchronous and spatially sparse and no more dense and synchronous as in traditional cameras, meaning that standard algorithms designed for image sequences are no more directly applicable. In contrast to traditional cameras which acquire information on absolute intensities, coping with binary events expressing only relative changes in light intensity is a non negligible challenge for standard applications. Moreover, because of the inherent shot noise in photons and the transistor circuit noise, all event sensors are noisy for their non-idealities [19], and the process of quantization of the light intensity changes is complex and has not been completely characterized.
However, many applications in different fields have been investigated throughout the last decades: feature detection and tracking, optical flow estimation, 3D reconstruction in both monocular and stereo, pose estimation and SLAM and image reconstruction. A more complete description on most of the event-sensing applications can be found at [20]. The most interesting applications for this work could be the different methods have been developed during the last decade for ray-scale image reconstruction from events. Due to the nature of the event-sensing cameras, some of these methods require also gray-scale offset images in order to recover the absolute brightness of the scenes, such as [21]. Nevertheless, some works [22], [23], [24], have used spatial and/or temporal smoothing in order to reconstruct the absolute brightness on the reconstructed gray-scale images without any knowledge of the initial gray-scale offset image. More recent approaches such as [25], [26], [27], [28], have instead developed *deep learning* algorithms for real time reconstruction of image frames starting from events only. By

reconstructing the gray-scale images from the events information it is possible then to apply standard algorithms for image processing. However, since the quality of the reconstructed images is directly affected by noise due to the non-idealities of event-cameras [19], as possible to see from the state-of-the art *FireNet* [27] grayscale-image reconstruction proposed in figure 16, and since the target of this work is to pixel-wise reconstruct the TRS from events in order to then reconstruct the surface normals on a scene, already existing methods for reconstructing gray-scale images are not taken in consideration.

(a) Ground-truth gray-scale image. (b) Reconstructed gray-scale image.

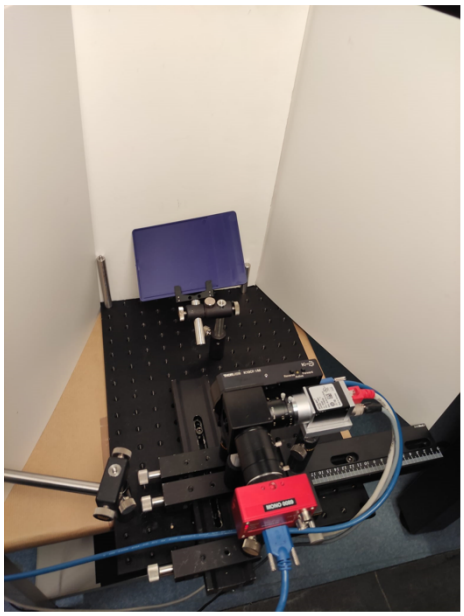
Figure 16: FireNet result on gray-scale image reconstruction from event camera output. [27]

4 Laboratory Setup

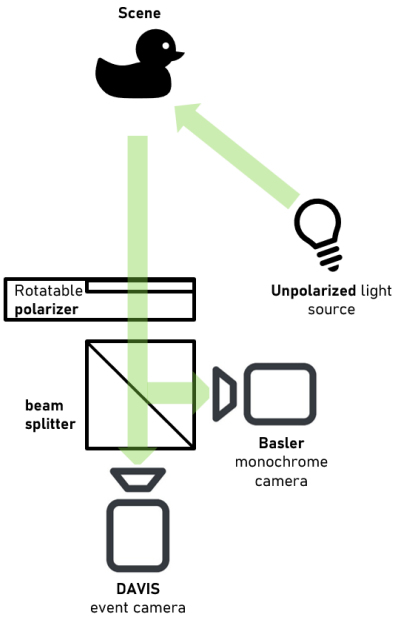
This chapter presents the full laboratory setup used for collecting data, then all the devices are presented separately by also emphasizing the parameters used for each device for the data collection phase.

4.1 Full Laboratory Setup

Since the aim of this work is to evaluate the possibility to perform surface normal reconstruction with polarization cues and event-sensing camera rather than traditional monochrome cameras, the devices used for this experiment are the DAVIS 346 RED event-sensing camera, a monochrome camera manufactured by *Basler*, a rotation stage in which a polarization filter is mounted on and a non-polarizing beam splitter. A more detailed description of the devices will be presented in the following subsections. Figure 17 graphically illustrates the final laboratory setup, with a close up view of the main devices shown in Figure 18.



(a) Picture of the full setup used for the data acquisition phase.



(b) Synthetic illustration of the setup. An example of light beam depicted in green.

Figure 17: Full setup used for the data acquisition.

The target of using a deep learning approach for pixel-wise reconstructing the signal of the event-camera as if it was acquired from a standard monochrome camera leads to

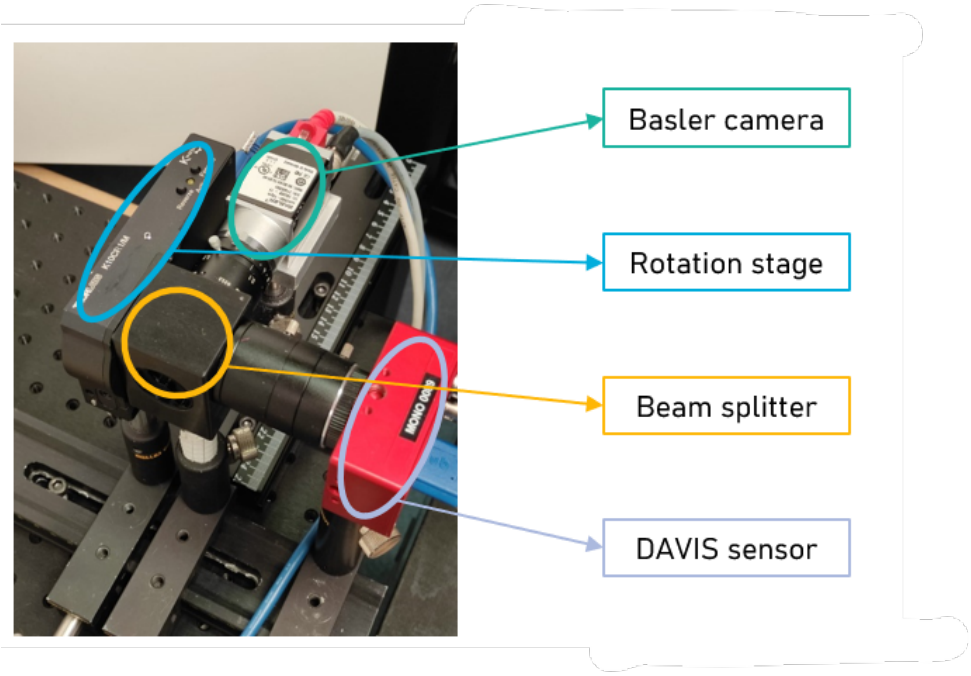


Figure 18: Close up of the devices used in the laboratory.

the necessity of introducing a $50R/50T$ non-polarizing beam splitter, which is an optical device that allows to split the input light into two perpendicular and equal beams having both 50% of the initial input intensity and the same polarization state of the input light. By placing the beam splitter in between the two cameras it is possible to capture the same scene as if the two cameras were placed on the same point in space and without changing the polarization angle of the input light, as the schematic of Figure 17b illustrates. This setup allows to have ground-truth data from the polarization images collected with the monochrome camera, allowing to train the neural network presented in section 6.2 in a fully supervised way. For matching as much as possible the field of view of the two cameras in order to capture the same portion of the scenes, different camera lenses available in the laboratory have been combined to achieve a satisfactory result. However, due to the method used for training the neural network, the homography computation for aligning the two cameras for a pixel-to-pixel match has been avoided.

It is important to notice that the initial setup did not take in consideration the white panels placed behind the scenes, since the initial target was to perform surface normal reconstruction by mainly exploiting diffuse polarization as in [4]. These panels were added lately in the project due to the difficulty of collecting events related only to polarization by diffuse reflection of the the objects in the scenes. As illustrated in Figure 19, with

the white panels it is possible to exploit the polarization by specular reflection which, having an higher DOP compared to diffuse reflection, has allowed to produce events on the DAVIS camera while rotating the polarization filter placed on the rotation stage. However, even if this way allow to collect events related to polarization information, this method has the disadvantage of the impossibility to disambiguate the Zenith angle of the surface normals due to the domination of the specular reflections on the diffuse ones and due to the single-view and fixed light source position of the setup as discussed in subsection 2.2.

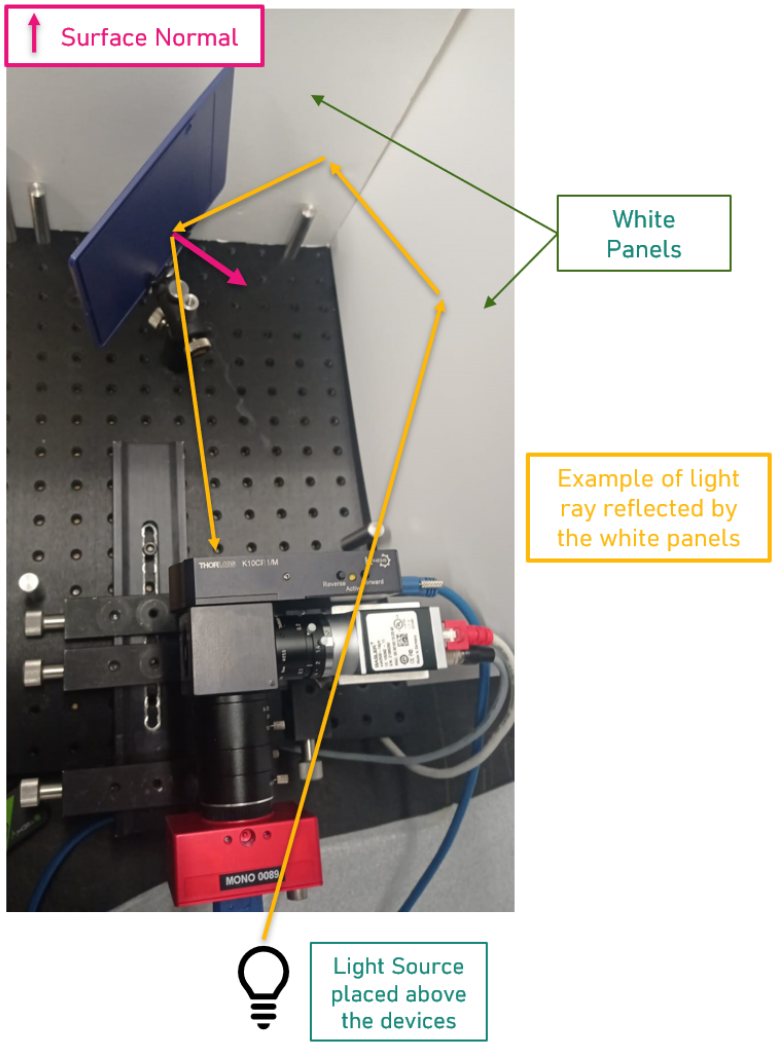


Figure 19: Illustration of specular reflection exploited by white panel placed behind the collected scenes. The higher DOP of the polarization by specular reflection has allowed to collect events related to polarization cues.

Furthermore, before the data acquisition, the polarized filter has been calibrated with a *polarimeter* in order to determine the polarization direction of the linear polarizing filter, which is an essential step in order to correctly recover the normals of the surfaces from polarization by reflection.
Regarding the light source, an halogen *Esser Test-Charts* illuminator of Figure 20 was used in order to produce a smooth and diffuse light, with the advantage that the light intensity could be modified in order to test in particular the behavior of the DAVIS sensor.

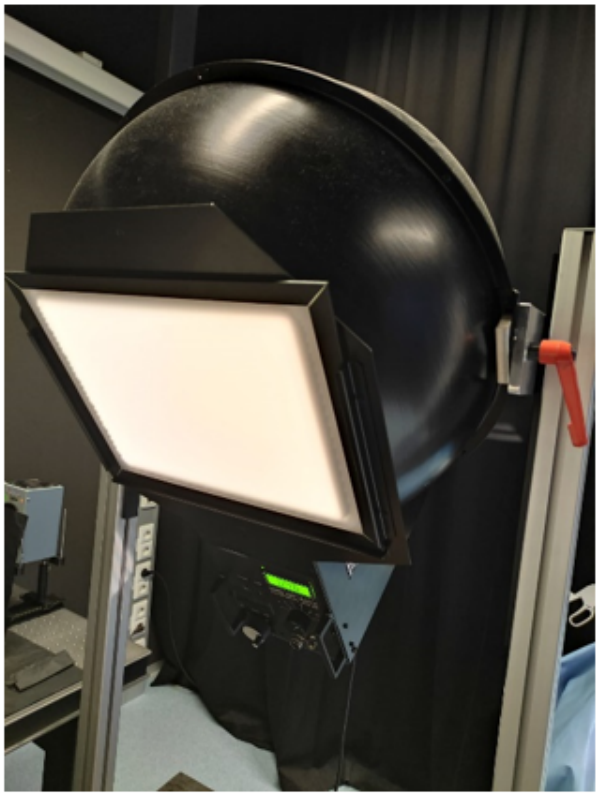
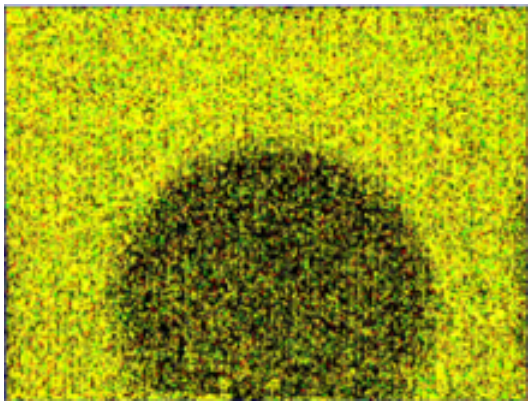


Figure 20: Light source used for the experiment (Esser Test-Charts Illuminator) in order to produce a smooth and diffuse light.

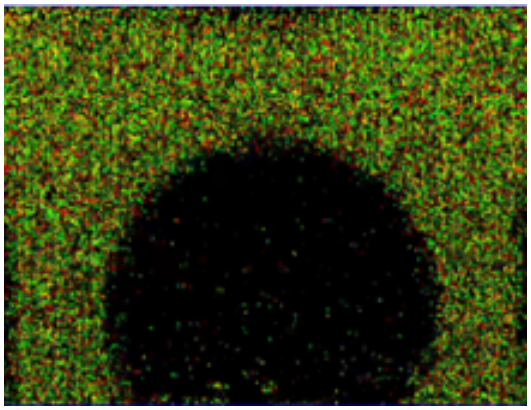
4.2 DAVIS 346-Red Event camera

A long time has been spent for setting up the DAVIS 346-Red event-camera for data acquisition. The main struggling point was to make the DAVIS spike events related to diffuse polarization in order to perform surface normal reconstruction by avoiding the Zenith angle ambiguity. For this purpose, many tests have been dedicated to modify the parameters of the DAVIS camera and evaluate its behavior during a complete

$360°$ rotation of the polarization filter in order to find a satisfactory combination of the parameters.
The event-camera behavior has been evaluated both visually, with the jAER software, and also by analyzing through Python plots the correlation between the pixels ON/OFF events and the expected TRS computed from the monochrome camera images. The main DAVIS parameters that have been modified are the ON-Threshold, OFF-Threshold and the refractory period. After numerous trials with different combinations of these parameters, the introduction of the white panels was required in order to increase the DOP of the reflected light and so to make the DAVIS sensor produce events related to the TRS while rotating the polarization filter. In fact, the low variation of intensity of the TRS for diffuse polarization pixels leads to the necessity of decreasing both ON and OFF event thresholds with respect to the values set by the sensor's manufacturer, however the threshold reduction makes the DAVIS pixels produce mostly noisy events even in static scenes where no events should be outputted, as in Figure 21a, where the red, yellow and green dots indicates events spiked by the corresponding pixel.



(a) DAVIS static frame with lower ON-OFF thresholds with respect to nominal threshold values set by the manufacturer, and nominal refractory time.



(b) DAVIS static frame with same ON-OFF thresholds as in the Figure 21a, but with lower refractory period with respect to nominal value set by the manufacturer.

Figure 21: Static scenes collected from DAVIS camera. Red, yellow and green dots corresponds to events outputted by the pixels. The scene captured by the sensor is on the darker area of the frame with circular shape, while the rest of the frame corresponds to the cage of the beam splitter where events are spiked due to shot-noise of the sensor.

In order so to be able to set lower ON-OFF thresholds, the refractory time of the DAVIS sensor is then decreased with respect to the nominal value in order to still being able

to capture events related to polarization but with a reduced the sensor noise due to the lower firing rate of the pixels, as it is possible to visually evaluate by comparing the noise on the static scene of Figure 21b with noise of the same scene in Figure 6a. The final parameters used for the DAVIS camera and setted in the jAER software for the data acquisition with the event camera are reported in Table 1

DAVIS sensor Parameters	
ON - Threshold	$18.9\,\%$
OFF - Threshold	$-14.2\,\%$
Refractory Time	$53,4\,\mu s$

Table 1

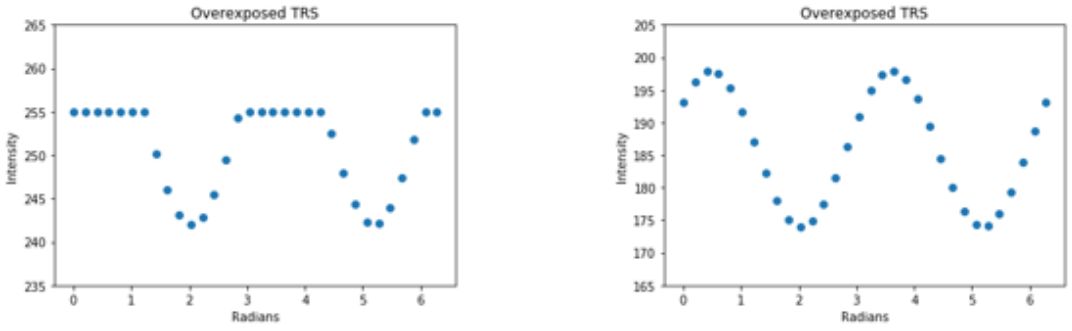
4.3 Monochrome camera

The monochrome camera used for this work is a Basler *acA2500-14gm*, and the main specifications are reported in Table 2:

Basler Camera Specifications	
Sensor Size (H×V)	$5.7\,mm \times 4.3\,mm$
Resolution (H×V)	$2592\,px \times 1944\,px$
Pixel Size (H×V)	$2.2\,\mu m \times 2.2\,\mu m$
Frame Rate	$14 fps$
Sensor Type	*CMOS*

Table 2

The full resolution of the camera is used for the data collection and the exposure time of the camera is set to $44975\,\mu s$ in order to avoid having over-exposed samples in the transmitted radiance sinusoids and so to preserve the polarization information of the light reflected by the objects in the scenes as in Figure 22.
For this purpose, by fixing the light source intensity, the aperture of the lens and the exposure time of the monochrome camera, a set of polarization images of a given scene



(a) Example of wrong exposure time.

(b) Example of correct exposure time.

Figure 22: Graphical examples of loss of information in the TRS samples when Basler camera is overexposed.

is captured at different polarization filter angles, then the TRS of the brightest pixels on the set of polarization images are evaluated in order to obtain a sinusoid-shape intensity in function of the polarization filter angle, as in Figure 22b. If the evaluated TRS instead has intensities clipped to 255 at some polarization angles, due to the $8bit$ intensity depth of the camera, then this procedure is repeated again by reducing the exposure time until a satisfactory result is reached.

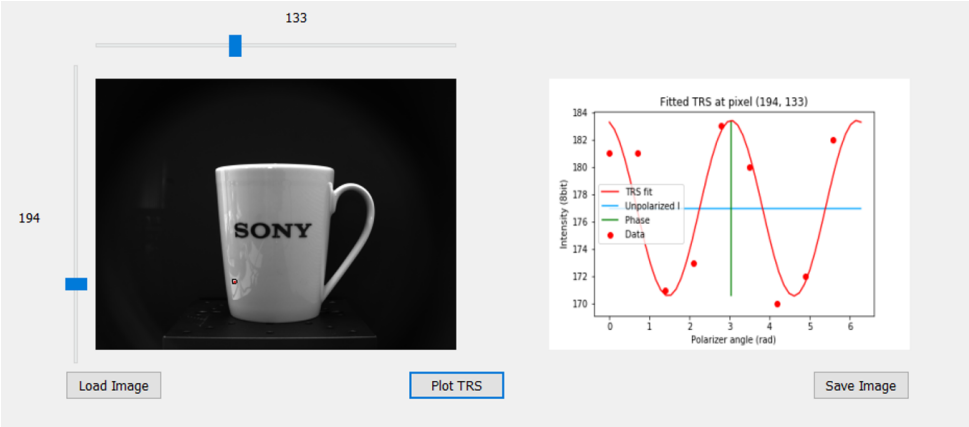


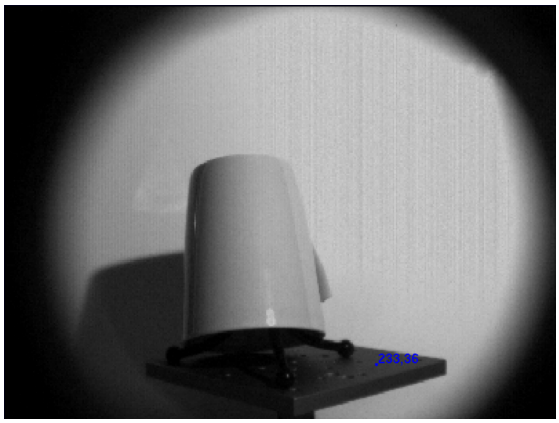
Figure 23: Graphical User Interface implemented for a fast pixel-wise evaluation of the TRS shape on a captured set on polarization images.

Moreover, a graphical user interface (*GUI*), briefly illustrated in Figure 23, has also been implemented in Python programming language for a fast pixel-wise evaluation of the TRS samples captured at different polarization filter angles. In this way it is possible to pixel-wise scan each set of polarization images and then plot the TRS corresponding to the selected pixel, allowing to speed up the procedure of correctly setting the exposure time of the monochrome camera and also to evaluate the DOP on the different surfaces

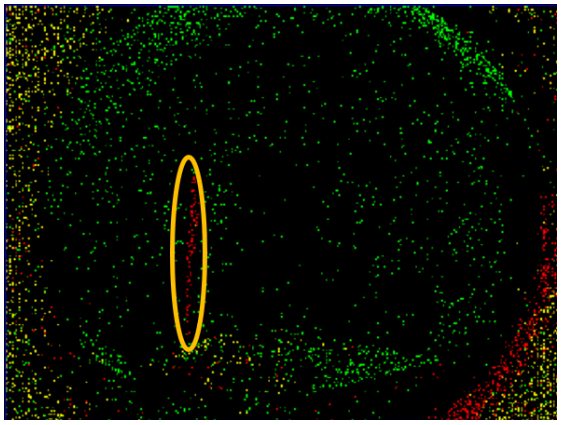
of the analyzed objects.

4.4 Rotation Stage

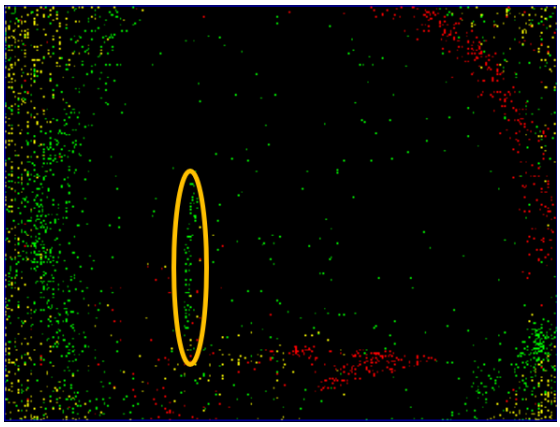
The *Thorlabs K10CR1/M* rotation stage used in the laboratory setup has a maximum angular velocity of $25°/s$, a maximum angular acceleration of $25°/s^2$ and a continuous travel range of $360°$. During some testing done with the data collected with the DAVIS camera through a continuous $360°$ rotation of the polarization filter, it has been noticed that the angular velocity of the rotation stage has a direct influence on the events produced by the sensor and related to polarization. In fact, by performing a slightly faster manual rotation of the polarization filter, it is possible to start seeing events related to reflection by diffuse polarization as in Figures 24b and 24c, which are not clearly present when rotating the rotation stage at its full velocity. So by visually comparing Figure 24b with 24d and Figure 24c with 24e, it is possible to deduce that the rotation velocity of the polarization filter could have a direct influence on the events spiked for light intensity changes related to polarization from diffuse light. This is probably due to the limited bandwidth of the DAVIS 346-Red sensor and since a full $360°$ rotation of the polarization filter takes around $15,4s$, considering the acceleration and deceleration phases, meaning that events should be produced duering the almost $4s$ it takes the TRS to go from its maximum value, to its minimum value and vice-versa, but being this time quite long and the TRA amplitude small for diffusely polarized light, events may be lost by the camera.
However, for in final setup used for acquiring data, the Thorlabs K10CR1/M rotation stage is used at the maximum angular velocity and acceleration available.



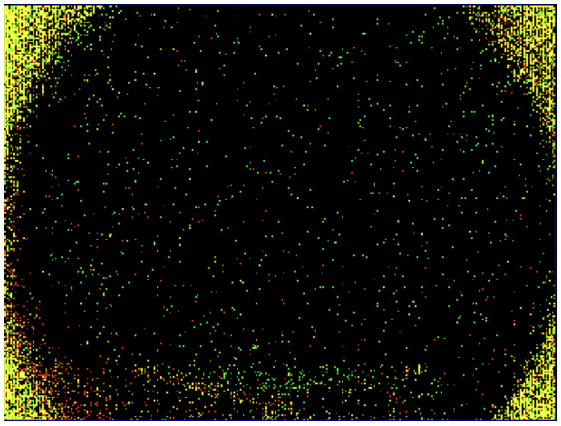
(a) Scene captured with APS sensor of DAVIS.



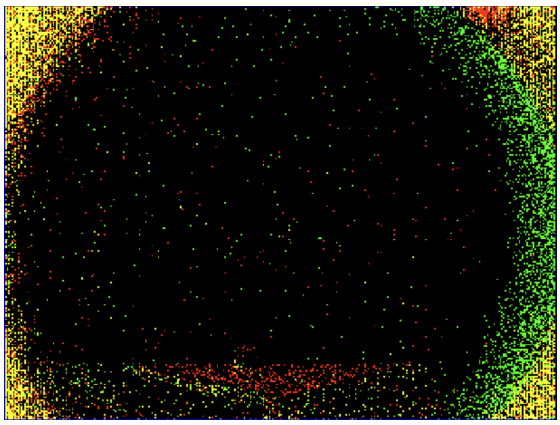
(b) Single events frame from manual polarization filter rotation. Circled in yellow, OFF events on the mug limb related to a descent from the TRS peak.



(c) Single events frame from manual polarization filter rotation. Circled in yellow, ON events on the mug limb related to an ascent towards the TRS peak.



(d) Single events frame from electronic polarization filter rotation. We expect to have a situation similar to the Figure 24b



(e) Single events frame from electronic polarization filter rotation. We expect to have a situation similar to the Figure 24c.

Figure 24: Visual comparison of DAVIS events with manual and electronic rotation of the polarization filter.

5 Dataset Creation

5.1 Data Acquisition

The data acquisition from each single scene is composed by two distinct steps:

1. polarization images acquisition with Basler camera at discrete polarization filter angles multiple of 30° on a 360° rotation, for a total of 12 polarization images for each scene;

2. events acquisition with DAVIS sensor during a continuous 360° rotation of the polarization filter.

The data acquisition from each single scene is automatized through a Python script in order to automatically acquire data from both the Basler camera and the DAVIS sensor and then save the raw corresponding data. Ground-truth images are acquired with the first part of the acquisition, instead the raw events collected during the second acquisition step are used, after being processed, as input to the neural network for the TRS reconstruction.
As discussed in subsection 2.2, only four monochrome polarization images captured at 0°, 45°, 90°, 135° angles of the polarization filter are needed in order to reconstruct the TRS and then the surface normal for each single pixel. However, in order to have a more reliable fit on each pixel's TRS samples of the ground-truth data, the total number of polarization images per scene is increased to 12 to produce a total of 12 TRS samples per pixel and so a more reliable ground-truth fit on the TRS samples for each pixel, leading to a more accurate estimation of the phase and the offset of the sinusoid.
For the purpose of investigating the possibility to pixel-wise reconstruct the TRS from DAVIS events, data for training the neural network is collected only from homogeneous materials, in particular from some *RAL-Plastics* [29] with specific colors as reported in Figure 25. By using only homogeneous materials in the scenes it is possible to consider only the data on a region of interest (*ROI*) on the two cameras and so avoid the pixel-to-pixel alignment between the two cameras. In this way, for each scene with a RAL-Plastic sample we can conside a ROI of ~ 3418 pixels for the DAVIS sensor and a squared ROI of 70×70 pixels for the monochrome camera.

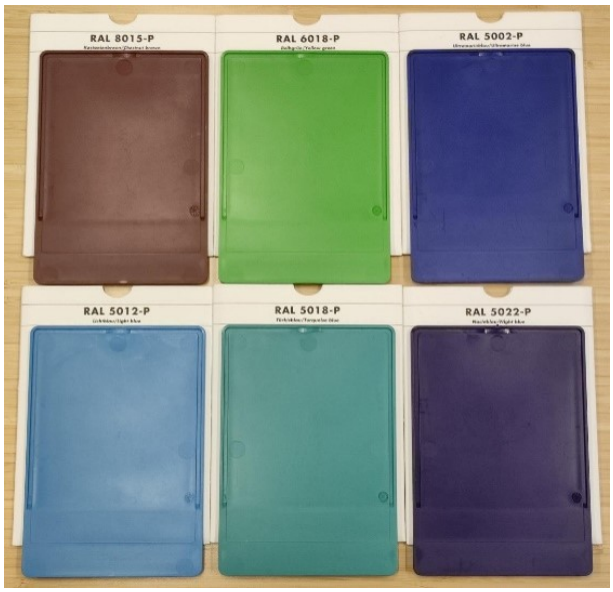
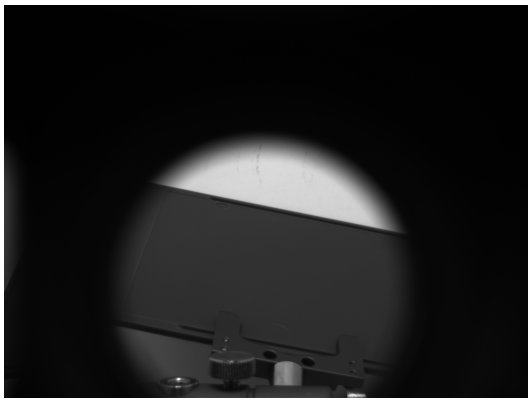
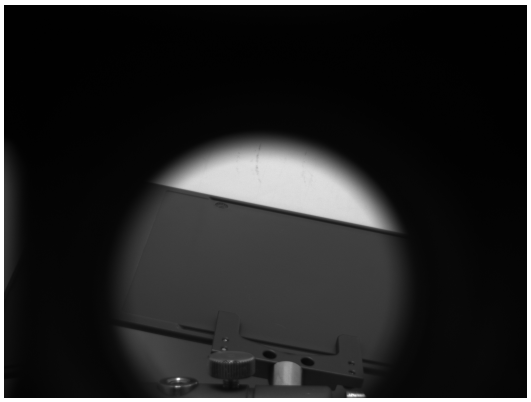


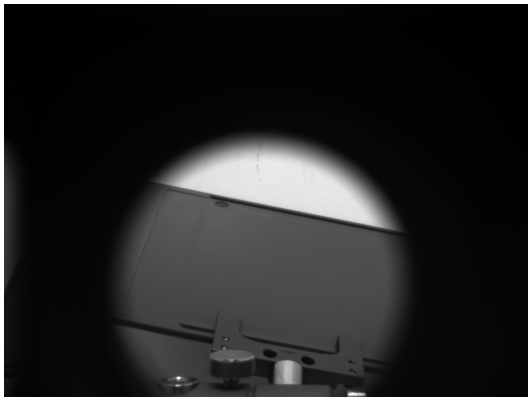
Figure 25: RAL-Plastic samples used for data collection.



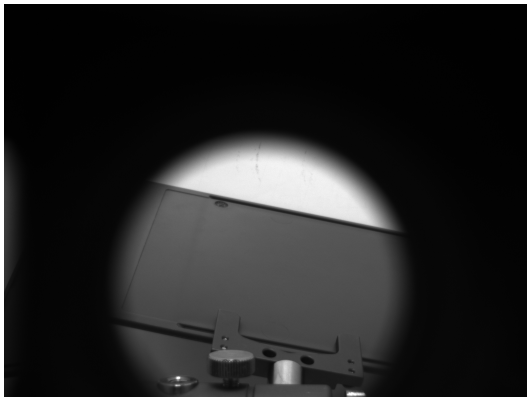
(a) Scene captured at $0°$ of the polarization filter.



(b) Scene captured at $30°$ of the polarization filter.



(c) Scene captured at $60°$ of the polarization filter.



(d) Scene captured at $90°$ of the polarization filter.

Figure 26: Examples of monochrome images captured at different polarization filter angles.

An example of monochrome images taken with the Basler camera at different polarization filter angles is shown in in Figure 26. Since the shape uniformity and the homogeneity of the RAL-Plastic, in the different points of the RAL-Plastic surface the intensity variation as the polarization filter is rotated is approximately the same. This means that on a full set of 12 polarization images, the TRS samples of the different pixels corresponding to the RAL-Plastic surface have approximately the same values. This is also visible from the Figures 26a and 26d, where the light intensity is uniform along the RAL-Plastic surface.

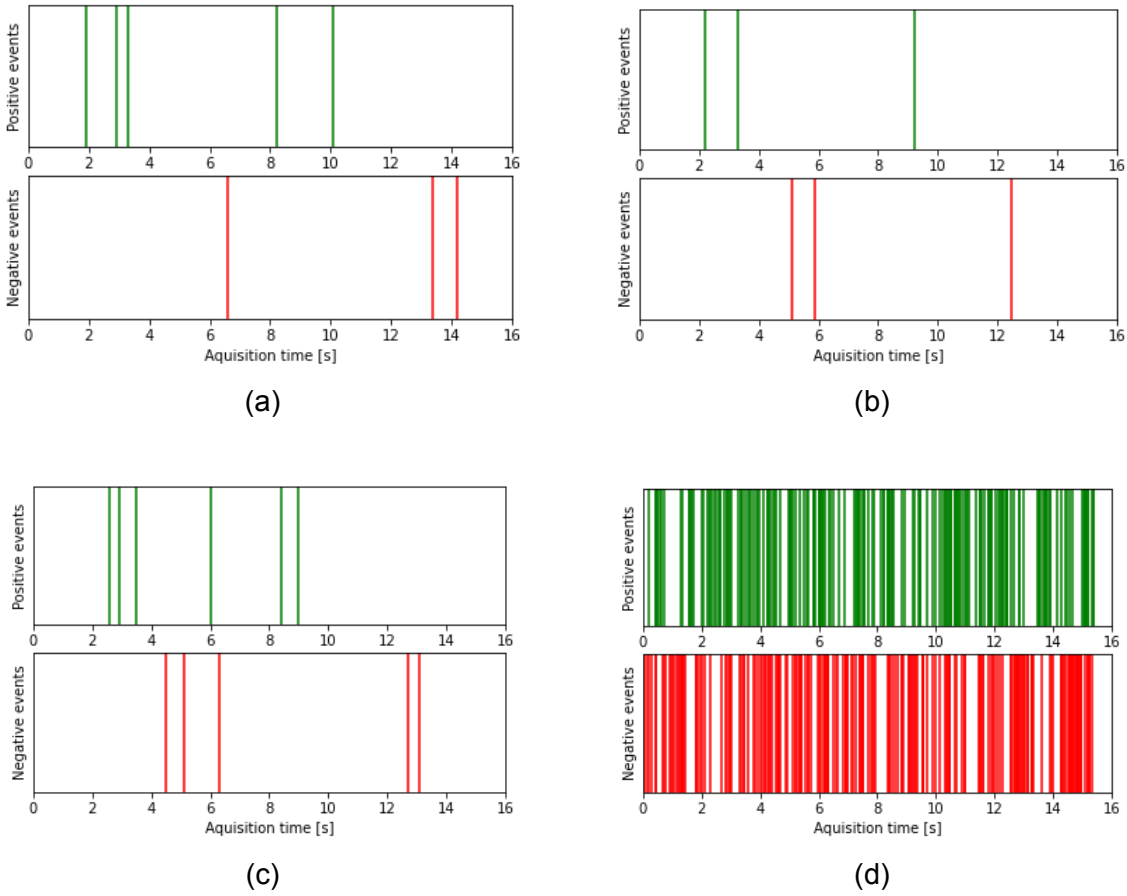


Figure 27: Plots of some different DAVIS pixels producing events over time as the polarization filter is continuously rotated for $360°$.

By considering the scene of Figure 26 as reference, the plots of the raw events produced over time by some different DAVIS pixels as the polarization filter is continuously rotated for $360°$ are presented in Figure 27. It should be noted that all the plots in Figure 27 are concerning DAVIS pixel capturing the surface of the RAL-Plastic sample of Figure 26, and the plot 27d represents a so-called *hot-pixel*, which are noisy DAVIS pixels

continuously producing ON and OFF events that are independent on light intensity changes [19].
Finally, to give a more general idea about the scenes from which the data is collected Figure 28 shows some examples of scenes acquired with the Basler camera.

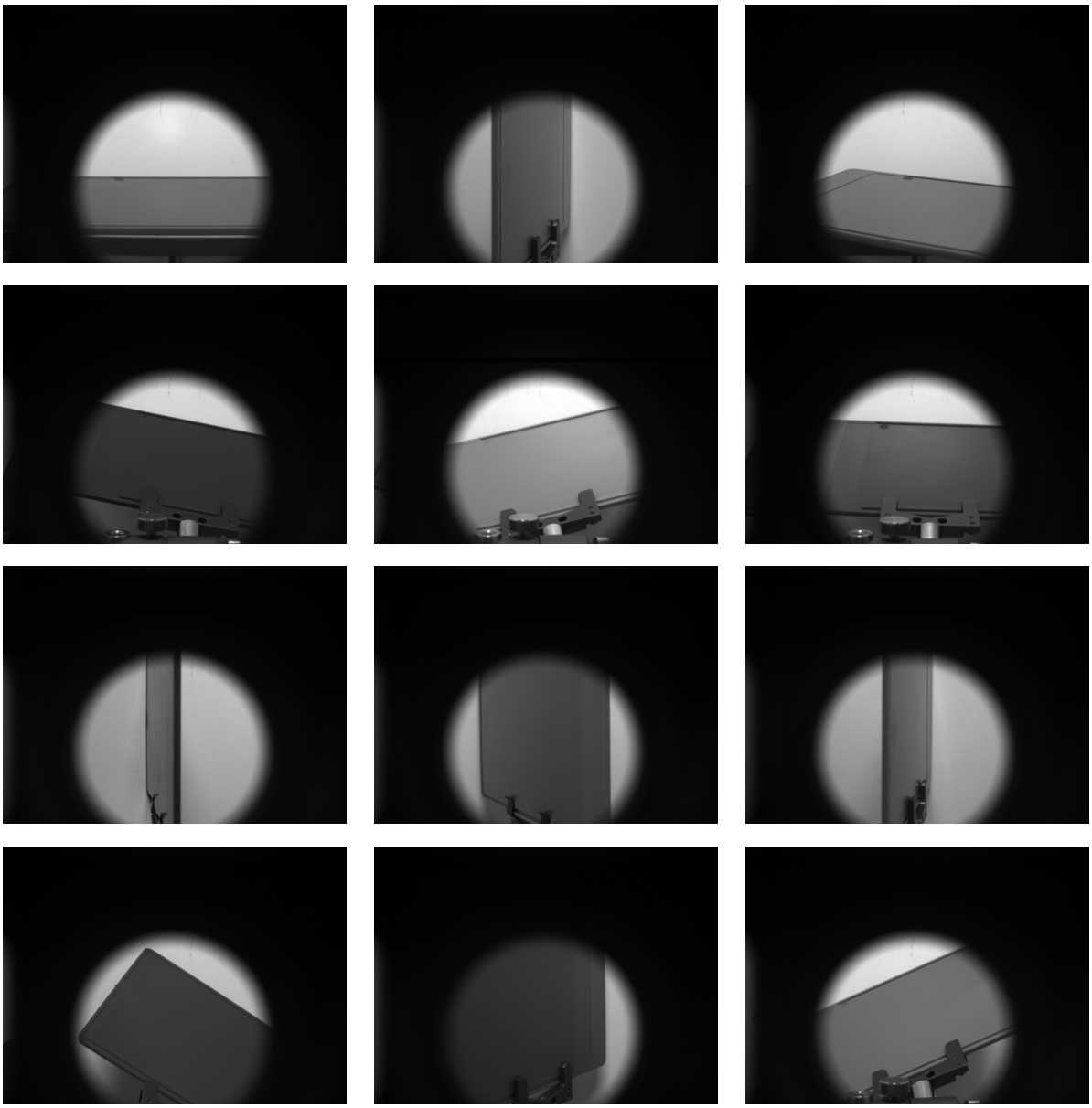


Figure 28: Examples of scenes acquired for the dataset creation. The images shown are taken with the Basler monoochrome camera.

To each scene on Figure 28 corresponds a unique surface normal associated to the RAL-plastic samples placed at different positions. However, the positions at which the RAL samples are placed are limited, as it will be discussed in 5.3. The total number

of scenes collected amounts to 105, and for each of them 12 polarization images are captured with Basler camera and then DAVIS events are collected during a continuous $360°$ rotation of the polarization filter.

5.2 Dataset Preparation

As discussed in the previous section, to avoid pixel-to-pixel alignment between DAVIS event-sensor and Basler camera, a ROI approach is used for creating the dataset and training the neural network. From the whole data captured from a scene, only the information on the two cameras ROI is then extracted for creating the dataset final dataset. Due to the non existence of a state-of-the-art neural network capable of performing a pixel-wise reconstruction of the TRS starting from event-sensing data, some different neural networks have been tested and the best results are achieved with a *1D-Unet*. The 1D-Unet takes as input the pre-processed events collected during a continuous $360°$ rotation of the polarization filter for a given DAVIS pixel, and outputs a total of 32 TRS samples, which are then compared, during the training, with the ground-truth TRS computed from the monochrome camera data.

5.2.1 Ground-truth Data Preparation

The data extracted from the polarization images captured with the monochrome camera are used as ground-truth data for a fully supervised neural network training. Due to the homogeneity of the RAL-Plastic samples captured in the scenes, on each set of 12 monochrome polarization images stacked together, the TRS samples are extracted only from a 70×70 ROI capturing the RAL-Plastic surface. Then all the TRS samples of the ROI are averaged together in order to produce a single TRS corresponding to the RAL-Plastic surface on the ROI. Thus, from each polarization scene only a single TRS is extracted, for a total of 105 ground-truth TRSs having different phase and offset. Then a sinusoidal fit on the averaged TRS is performed from which then a certain number of samples is extracted as ground-truth data for the 1D-Unet, as discussed in Chapter 6. Instead, as ground truth-data for training different networks that have been tested, after the sinusoidal fit on the averaged TRS samples, only the phase, the offset and the amplitude of the averaged TRS are extracted as ground-truth data. Figure 29 briefly illustrates the pipeline for the groud-truth data creation for the 1D-Unet training, consisting on a single TRS normalized in the range $[0, 1]$.
The final ground-truth dataset for training the 1D-Unet consists so of a matrix of shape

$(105, n)$, where n represents the number of samples considered from the averaged TRS and has been chosen during the training phase.

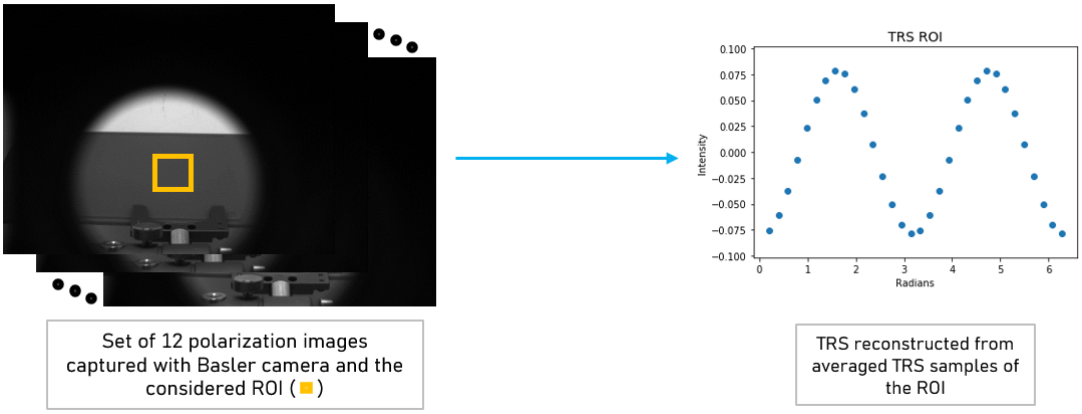


Figure 29: Brief illustration of data extracted from a set of Basler polarization images for the ground-truth dataset creation.

5.2.2 Input Data Preparation

The input data for each tested network consists on the events collected on the DAVIS ROI selected from each scene. For each pixel of the DAVIS ROI considered for a given scene, the events captured during a full rotation of the polarization filters are processed in the following way:

1. Removal of the events collected while the DAVIS camera starts acquiring data but the rotation filter still have to start rotating;
2. Removal of the noisy hot-pixels;
3. Binning in time domain of the acquired events, and then for each bin:
    - Computation of the sum of the event polarities;
    - Extraction of total number of positive events;
    - Extraction of total number of negative events.

The process of removing the events collected while the DAVIS camera starts acquiring data but the rotation filter still have to start rotating is related to the not negligible delay that has been noticed from when the DAVIS sensor is activated for outputting events and the rotation stage starts rotating the polarization filter. This leads to an initial transient

phase in which the DAVIS sensor is acquiring data but the rotation stage is not activated yet, so noisy events that may occur during the initial static scene captured with the event-sensor are removed. Moreover, the delayed synchronization between the DAVIS sensor and the rotation stage delays also, in the time domain, the events related to polarization, which can propagate to a slightly phase shift the reconstructed TRS. The hot-pixels as in Figure 27d are removed since considered for definition to be noisy pixel for which a correlation between outputted events and the light intensity variation does not exist. After a first step of noise removal from the DAVIS data, the main step for the input data creation consists in the binning of the events in the time domain. Different number of time-domain bins have been chosen for the data creation in order to evaluate the neural network training capability with different input sizes. The most important step is however in which features choose to extract from each bin. Initially only the sum of the polarities ($+1$ if positive event, -1 if negative event) has been considered as a feature, but since in a time-bin there is the possibility to have a null sum of polarities due to an equal number of positive and negative events, extracting also the total number of positive and negative events from each time-bin is a good idea to make the neural network learn to reconstruct the TRS from knowing more information on the events occurring on each time-bin. A brief illustration of the feature extraction step with 4 time bins is presented if Figure 30.

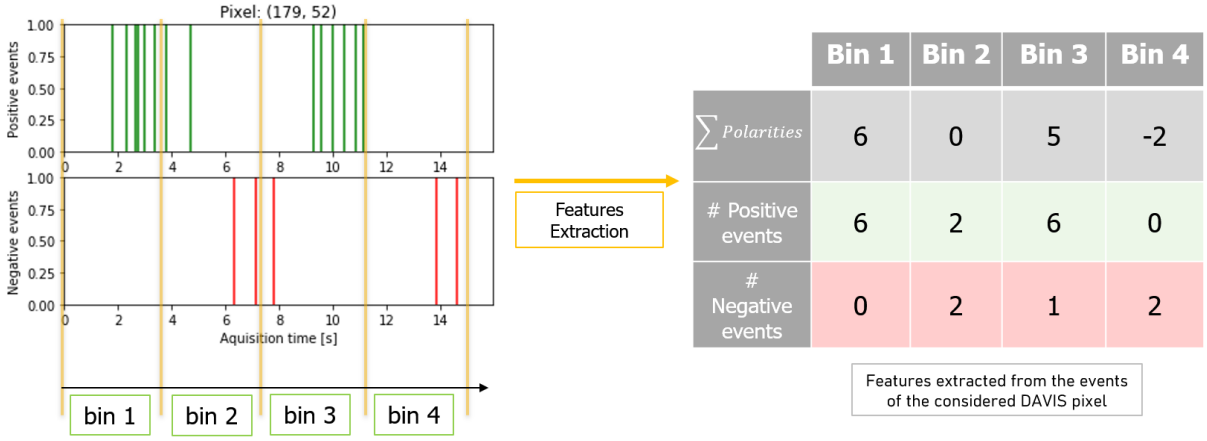


Figure 30: Brief illustration of feature extraction procedure from the events outputted over time by a single DAVIS pixel. In this example 4 time-bins are used for simplicity.

The final input dataset for training the 1D-Unet consists in a matrix of shape $(p, b, 3)$, where b determines the number of time bins, and $p = 358916$ represents the total number of DAVIS pixels considered in the ROIs of the overall collected scenes.

5.3 Dataset limitations

The reason why only a limited amount of RAL-Plastics is used for the data collection is because of the impossibility of acquiring events related to polarization for RAL-Plastics available in the laboratory and with color differing from the ones of Figure 25. In fact by fixing the devices settings, fixing the light source intensity and by placing RAL-Plastics with different colors on the exact same place on the scene, only for the RAL samples of Figure 25 the events produced by the DAVIS sensor can be related to intensity light changes due to polarization. This seems to be possibly related to the *Umov effect*, for which the degree of polarization of reflected light is depending also on the color of the objects reflecting the light, so on the wavelength reflected light. Thus, the lower degree of polarization for some RAL-Plastic samples rises the need to reduce the ON/OFF thresholds on the DAVIS camera. But, as previously discussed, the thresholds reduction leads to a a situation where the pixels mainly produce noisy events, making it impossible to reach a satisfying trade-off between events related to polarization and noisy events. Another limitation of the dataset is due to the orientation of surface normals collected in the scenes, so the orientation of the RAL-Plastic samples with respect to the cameras. By considering as reference the hemisphere of surface normals of Figure 31 and the partition of the hemispere in the regions A,B and C, the surface normals orientation acquired in the scenes for creating the dataset are belonging only to the region A.

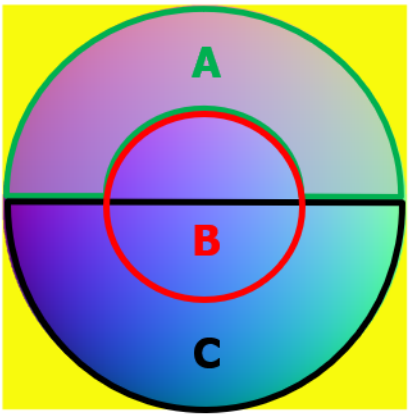


Figure 31: RGB-Encoded hemisphere of surface normals with the highlighted A,B,C regions. Surface normals orientation acquired in the scenes are belonging only to region A.

In the region B, where the surface normals orientation is pointing towards the viewer, the angle between the viewer and the surface normal is very low, thus the DOP on the

region B is very low as well, as discussed in subsection 2.1, making it impossible to acquire data with the DAVIS sensor. The impossibility of collecting polarization data for surface normals belonging to the boundaries of the region C instead is imposed by the laboratory setup used for acquiring data. This is due to the impossibility of making a unique triangularization between the devices position, the light source position and the white panels position for acquiring all the data by exploiting the reflections of the white panels as in Figure 19 and without having to change at least the position of the light source. However, acquiring data from multiple light source positions or multiple point of views of the scenes goes beyond the target of this work, as discussed in subsection 2.2.

6 Results

In this chapter are discussed the results on the training and TRS reconstruction performace of the tested neural networks. Furthermore, some results on the surface normal reconstructions by using the algorithm proposed by [4] are discussed in subsection 6.3. All neural networks tested are implemented with *TensorFlow*, moreover *Optuna* hyper-parameter optimization framework [30] is used for finding the optimal hyper-parameters for each network through different training trails.

6.1 First tests on the TRS reconstruction

Due to a lack in literature of of a deep learning algorithm for pixel-wise performing the TRS reconstruction from event-camera events, the first network tested for reconstructing the TRS starting from the binned events is a simple *1D convolutional neural network*

Figure 32: Multi-output CNN tested for reconstructing the Phase, DOP and Offset of the TRS.

(CNN) with a fully connected output layer, which takes as input the time-binned events of a DAVIS pixel with the feature extraction method presented in subsection 5.2.2, and outputs the phase, offset and DOP of the TRS to be reconstructed. From these output parameters in fact it is possible to fully reconstruct the TRS. However, this network does not produce any good results for the TRS parameter reconstruction, even for different hyper-parameters and depth of the CNN stack and the fully connected layer. For this reason, a multi-output CNN which is depicted for reference in Figure 32 has also been tested in order to output the phase, DOP and offset of the TRS by exploiting simultaneously different branches. In this way it is possible for each branch to set a specific architecture depth for the CNN layer and for the fully connected layer, and different hyper-parameters for each different branch. However, even the results for this method are not satisfying, reason why only the results with the 1D-Unet are discussed from now on.

6.2 1D-Unet for TRS reconstruction

By changing approach with respect to the output parameters of the network, another tested model is a 1D-Unet consisting of contracting and expanding paths, which is re-adapted from the 2D-Unet originally used for image segmentation task [31]. Figure 33 illustrates a block diagram of the 1D-Unet used for the TRS reconstruction task. This network takes as input the features extracted from the time-binning of the events for each DAVIS pixel, as explained in subsection 5.2.2, and outputs a certain number of samples of the TRS to be reconstructed rather than the phase, offset and DOP parameters of the TRS.

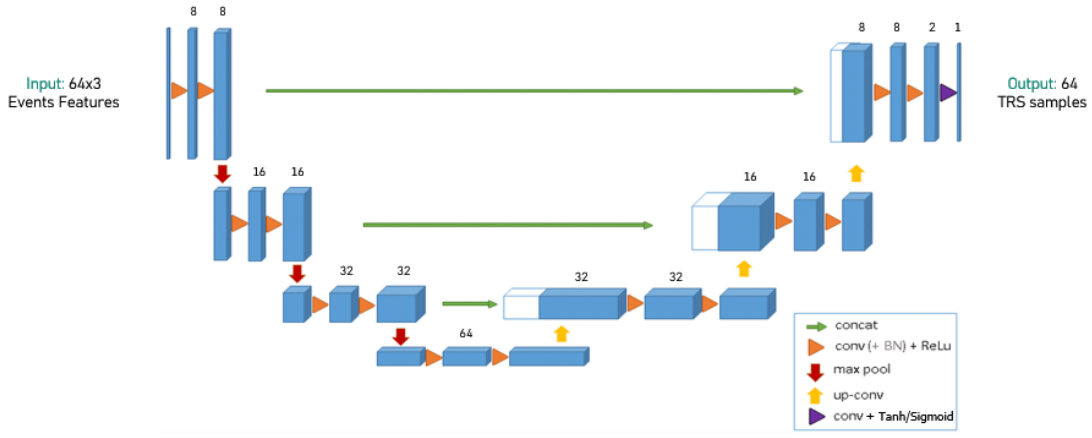


Figure 33: Scheme of the 1D-Unet model used for the TRS samples reconstruction.

Regarding the main architecture of the network, a 1D-Unet with 3 concatenation levels as in Figure 33 and a 1D-Unet with 4 concatenation levels are tested. For each of these architectures, Optuna framework is used in order to optimize different hyper-parameters: the input and the output shape, the optimizer, the learning rate and the momentum and finally the batch size of the input data. For each hyper-parameter optimization trial done by Optuna, the input shape $(n,\, 3)$ and the output shape $(n,\, 1)$ of the network are tested for $n = 16,\, 32,\, 64$. As optimizer, only *RMSProp*, *ADAM* and *SGD* are considered, each with initial learning rate belonging in the range $[0.00001, 0.1]$. The input batch size is instead chosen to be 32, 64, or 128. As training loss, the Mean Absolute Error (*MAE*) computed between the reconstructed TRS samples and the ground-truth TRS, is used. Overall, after a total number of 500 training trials performed with Optuna with different hyper-parameters, the lowest loss for the TRS samples reconstruction is achieved with the hyper-parameters of Table 3.

Hyper-paramater	**Optimal value**
Input shape	$(64,\, 3)$
Output TRS samples	64
Optimizer	*ADAM*
Batch size	32
Initial Learning Rate	0.00134
Loss	*MAE*

Table 3: Hyper-parameters used for the final training of the 1D-Unet.

As regularization technique for training the 1D-Unet, *early stopping* is used in order to stop the training phase if the validation loss does not decrease within 30 consecutive epochs. Batch normalization layers are added between each down-convolutional layer and max-pooling layer for a more stable training of the network. Moreover, a learning rate scheduler is used in order to halve the learning rate value every 15 training epochs as shown in Figure 34.

The 1D-Unet has been trained to reconstruct the TRS samples both with and without the offset, as it will be discussed next.

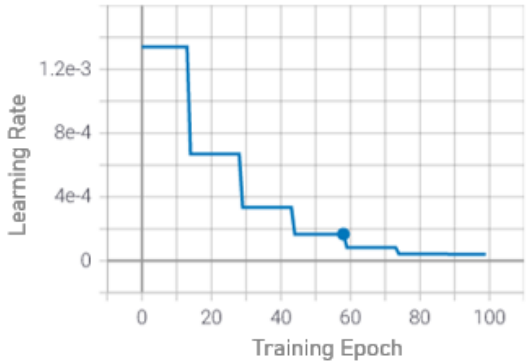


Figure 34: Learning rate values scheduled over training epochs for training the 1D-Unet.

6.2.1 TRS Rconstruction with Offset

The first trial of training the U-Net to reconstruct the TRS samples does not lead to good results. This is due to the non-capability of the network to learn to correctly detect the offset of the TRS by only exploiting the features extracted from the DAVIS events. In this case, the network is trained for a total of 100 epochs, the plot of the train and validation losses are presented in Figure 35 and the final scores for validation and test loss are reported in Table 4.

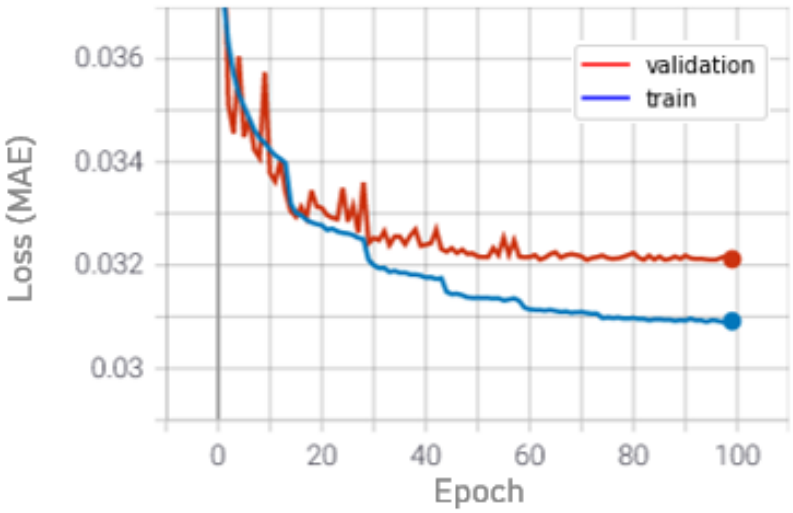


Figure 35: Train and validation losses (MAE) of the 1D-Unet with offset reconstruction.

Final loss scores on training 1D-Unet with offset	
Validation Loss	0.03277
Test Loss	0.03312
Number of test samples	35891

Table 4: Final loss scores for the 1D-Unet training with considering also the TRS offset reconstruction.

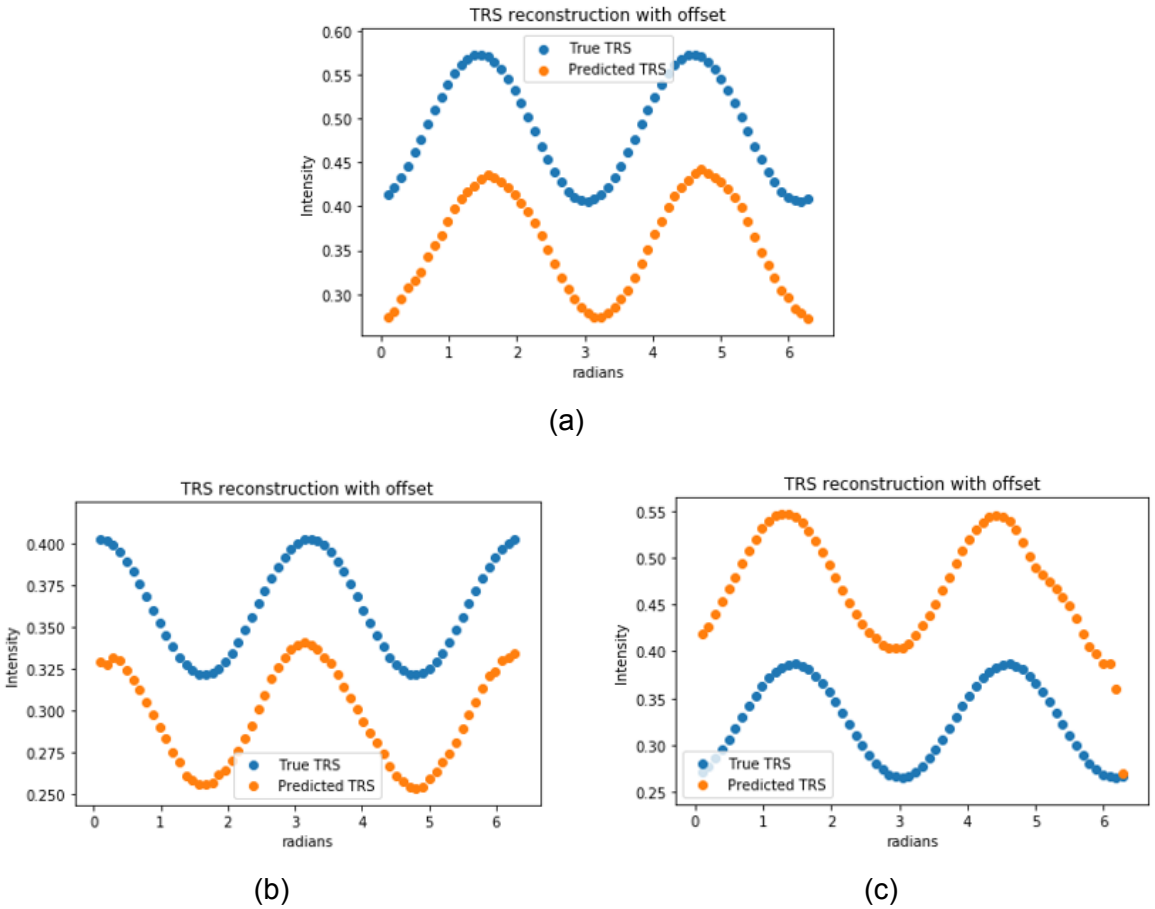


Figure 36: Examples of TRS with offset reconstructed after training the 1D-Unet. Blue dots represents the ground-truth TRS samples, orange dots represents the reconstructed TRS samples.

In Figure 36, some reconstructed TRS samples and some ground-truth TRS samples are plotted for a visual comparison. At a first glance it is evident that the main problem of this network is related to the offset reconstruction, reason why it has been decided to use the method presented in subsection 6.2.2, however the phases and the amplitudes of the sinusoids seems to be learned by the network. In surface normal reconstruction from polarization cues, a wrong TRS offset reconstruction leads to a wrong computation of the DOP and then to a wrong computation of the Zenith angle of the surface normal as, discussed in subsection 2.2.

6.2.2 TRS Rconstruction without Offset

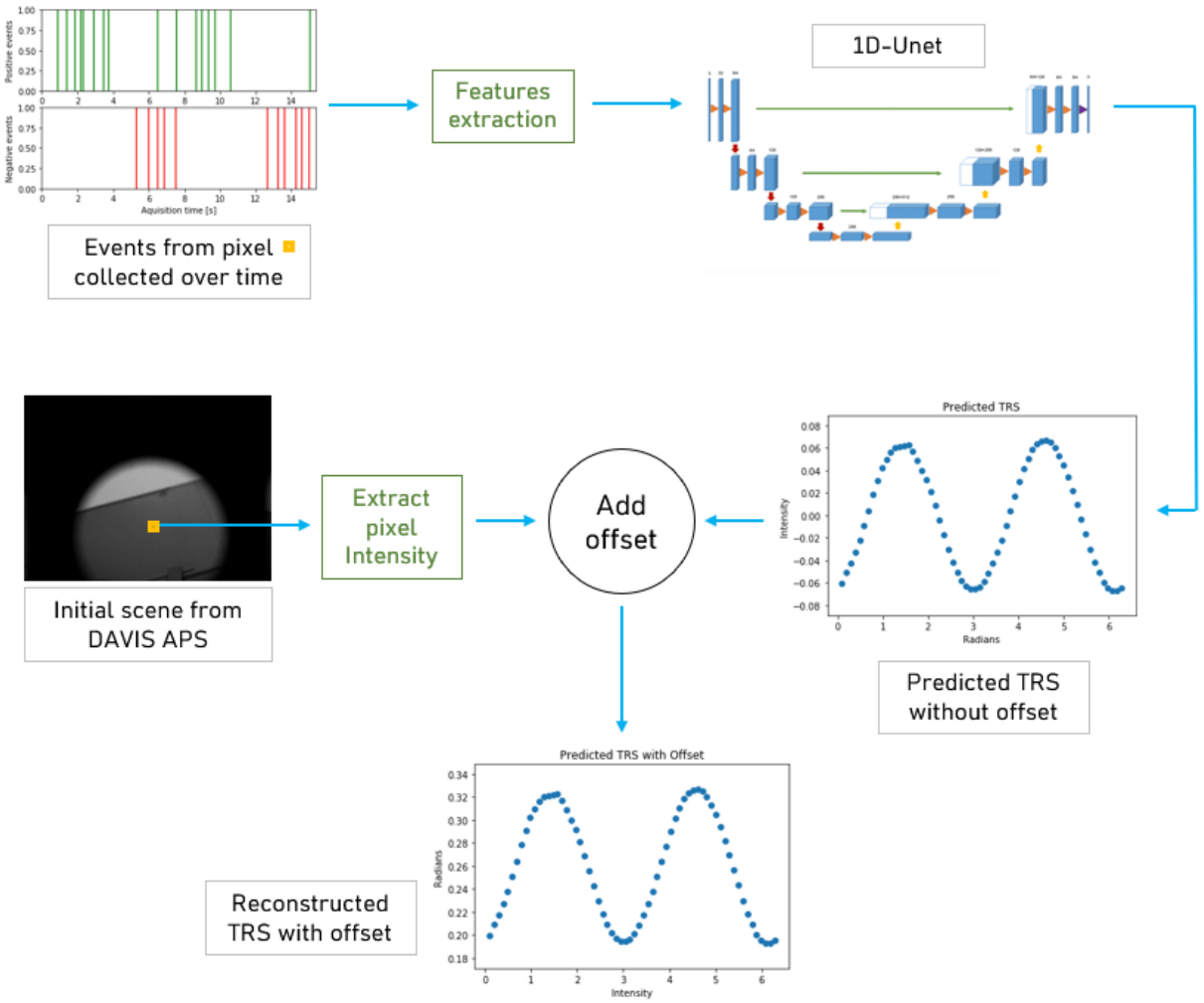


Figure 37: Schematic illustration of the method used for reconstructing the TRS offset by exploiting the DAVIS APS frames.

To overcome the difficulty of making the network to learn to reconstruct the offset of the TRS, it has been decided to exploit the APS circuitry of the DAVIS camera in order to capture a monochrome image with the DAVIS camera at $0°$ angle of the polarization filter, in order to extract the initial absolute intensity of each pixel from the scene before starting to acquire events while rotating the polarization filter. For this purpose, the exposure time of the DAVIS camera is set in order to match the pixels intensities of the scenes acquired with the Basler camera, in order then to have the same TRS offset from both cameras.

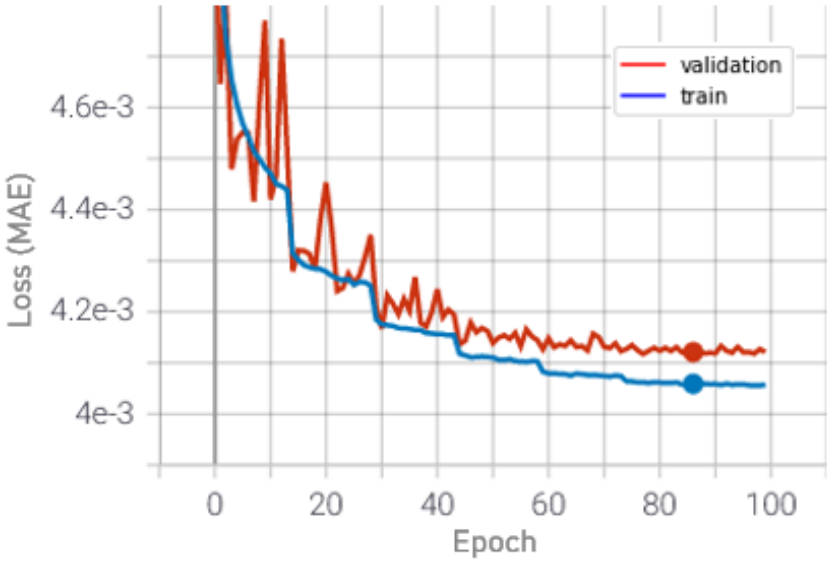


Figure 38: Train and validation losses (MAE) of the 1D-Unet without offset reconstruction.

Final loss scores on training 1D-Unet without offset	
Validation Loss	0.0041762
Test Loss	0.0041921
Number of test samples	35891

Table 5: Final loss scores for the 1D-Unet training with considering also the TRS offset reconstruction.

With this method it is possible to train the 1D-Unet to reconstruct for each DAVIS pixel the TRS without the offset and then add the offset extracted from the initial monochrome image of the scene, as illustraded in Figure 37. The validation and the training losses of

the final training of the 1D-Unet without the TRS offsets and with the hyper-parameters of Table 3 are reported in Figure 38. The final loss scores instead are reported in Table 5. By using this method, the final test loss is decreased of a facor of 10^{-1} when compared to the final test loss of the 1D-Unet trained for reconstructing the TRS with offset. Some final TRS reconstructions by using this approach are presented in Figure 39.

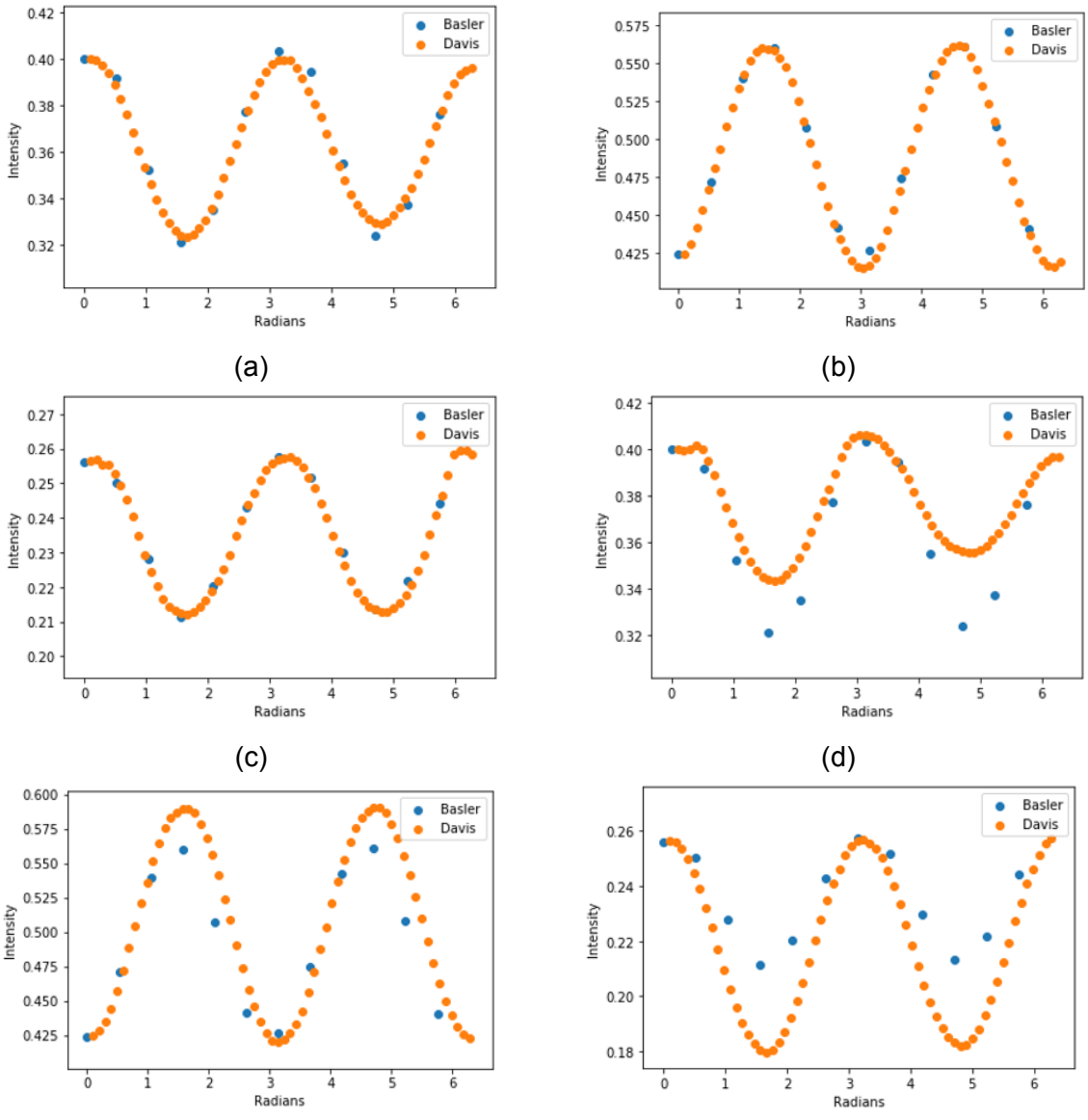


Figure 39: Examples of TRS reconstructed by exploiting the DAVIS APS for the TRS offset reconstruction. Blue dots represents the ground-truth TRS samples, orange dots represents the reconstructed TRS samples.

As will be discussed also in the results on the surface normals reconstruction of subsection 6.3, this method is still not perfect in the TRS reconstruction. In some cases, as the ones in Figures 39d and 39f the amplitude of the TRS is not correctly reconstructed by the 1D-Unet and this directly affects the surface normal computation by slightly distorting the predicted DOP. However, by reconstructing the offset by using this strategy, the errors on the TRS offset reconstruction are remarkably reduced as comparable between the plots if Figure 36 and the ones in 39. Moreover, as it is possible to visually evaluate from Figure 39e, beside the small difference in the predicted amplitude, the predicted phase is slightly shifted, and this will lead to some problems in reconstructing the surface normal orientations belonging to the *east* and *west* regions of the surface normals hemisphere of Figure 31, as it will be discussed in subsection 6.3.

6.3 Surface normal reconstructions

After reconstructing the TRS and extracting the parameters of the sinusoid, we are able then to reconstruct the surface normal for each pixel by using the method proposed in [4]. However, due to the impossibility of solving the Zenith angle ambiguity of the surface normals, since the data is collected by relying on polarization by specular reflection as discussed in subsection 5.1, the Zenith angle disambiguation is solved manually.
The surface normal reconstruction is discussed for three different scenes, by considering as reference the RGB surface normals encoded as in Figure 40.

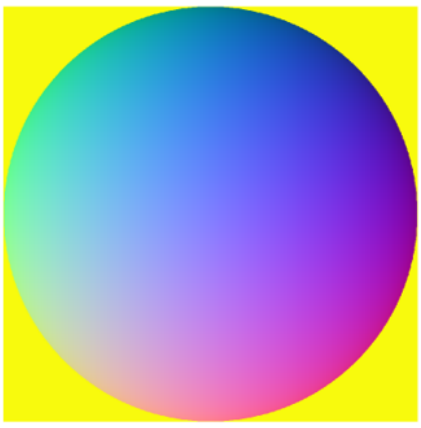
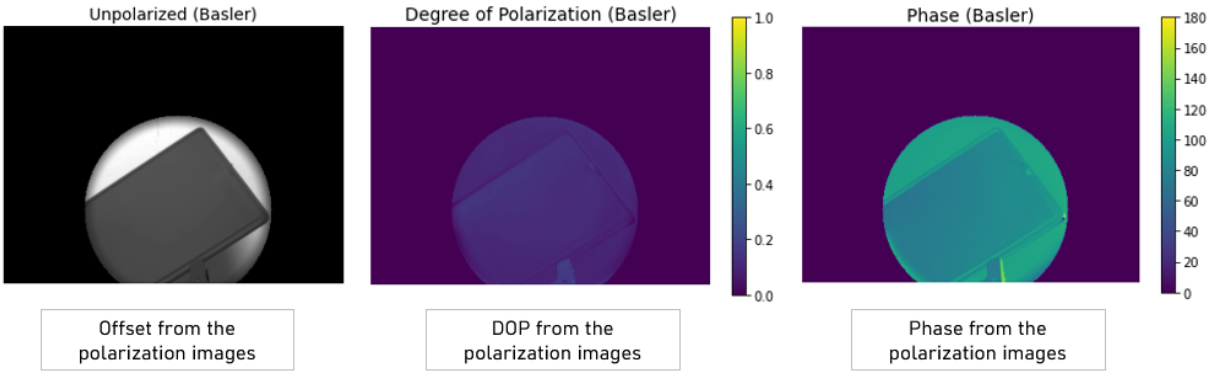
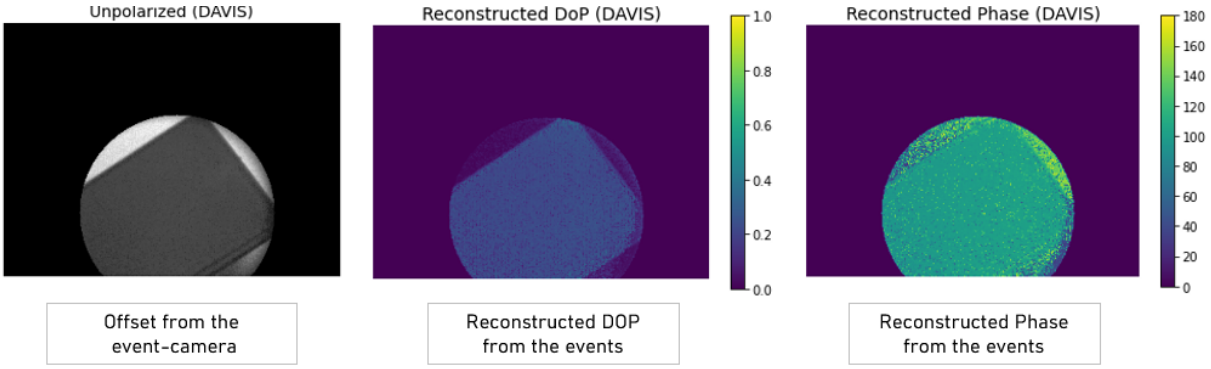


Figure 40: RGB-Encoded surface normals used for the normal reconstructions.



(a)



(b)

Figure 41: Image representation of the DOP and phase the from the Basler ground-truth TRS samples (a), and from the reconstructed TRS samples (b) with the method proposed in subsection 6.2.2. Scene A.

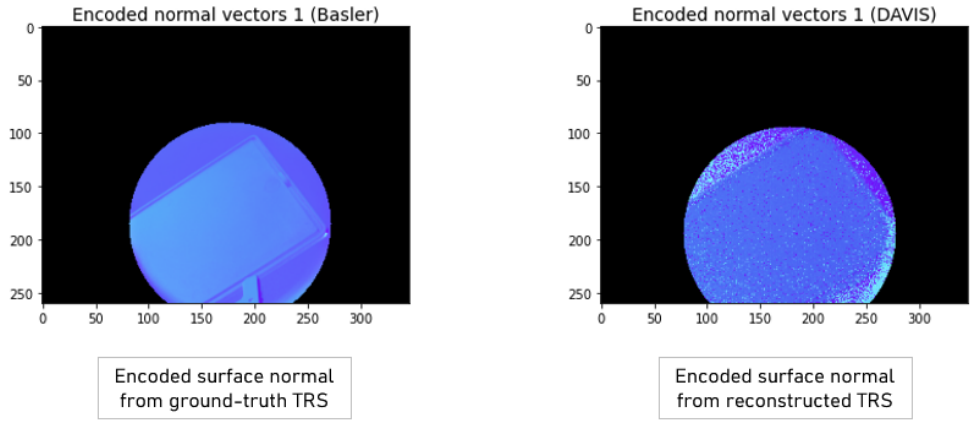
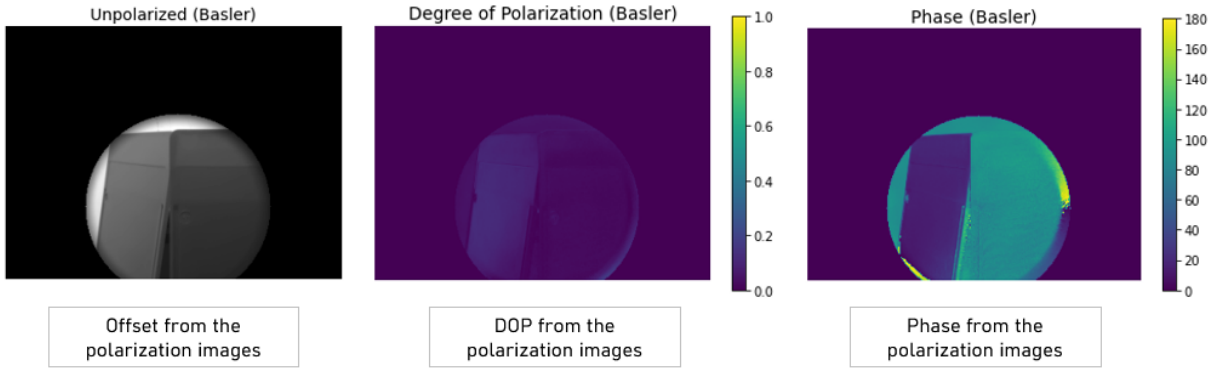
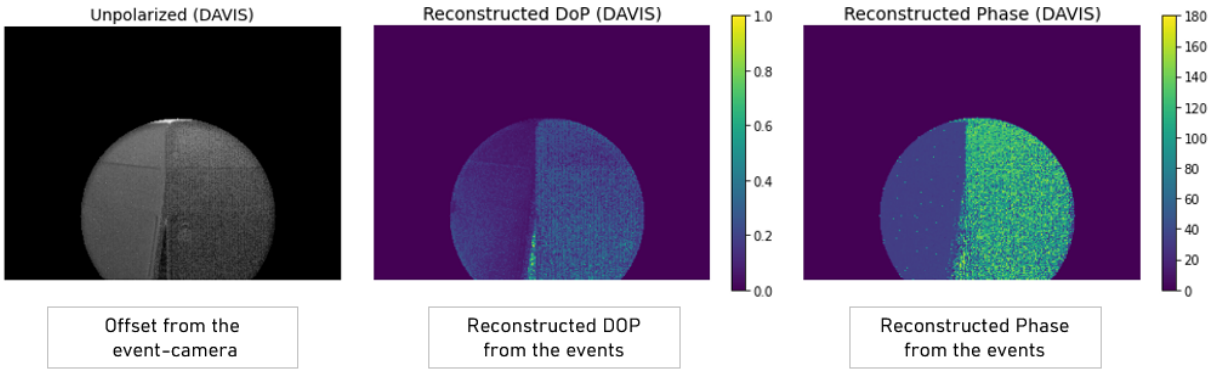


Figure 42: Encoded surface normal reconstructed from scene A.



(a)



(b)

Figure 43: Image representation of the DOP and phase the from the Basler ground-truth TRS samples (a), and from the reconstructed TRS samples (b) with the method proposed in subsection 6.2.2. Scene B.

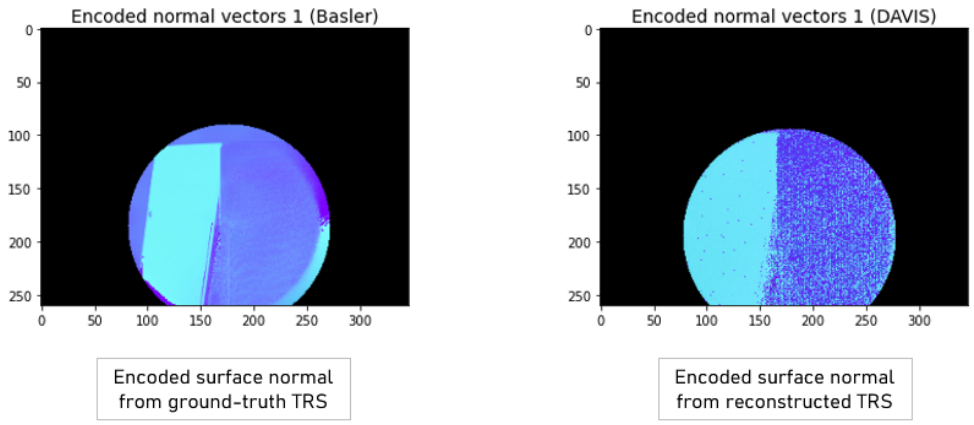
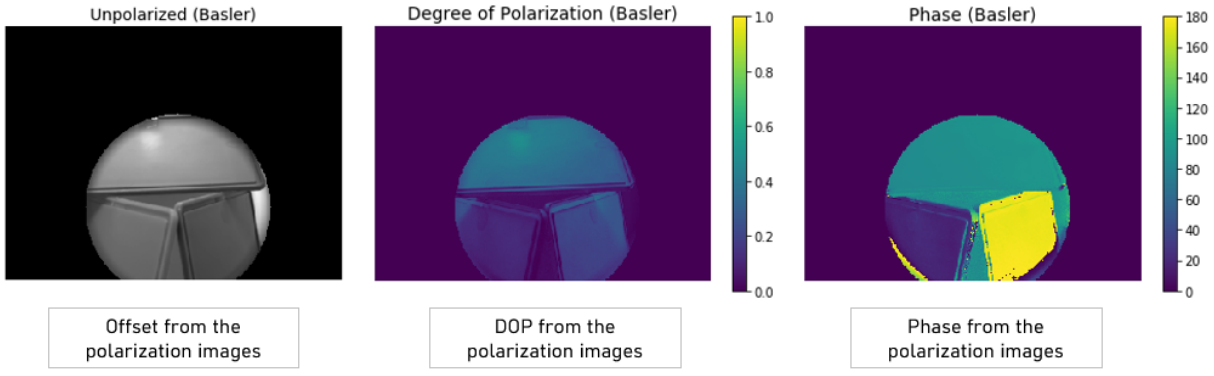
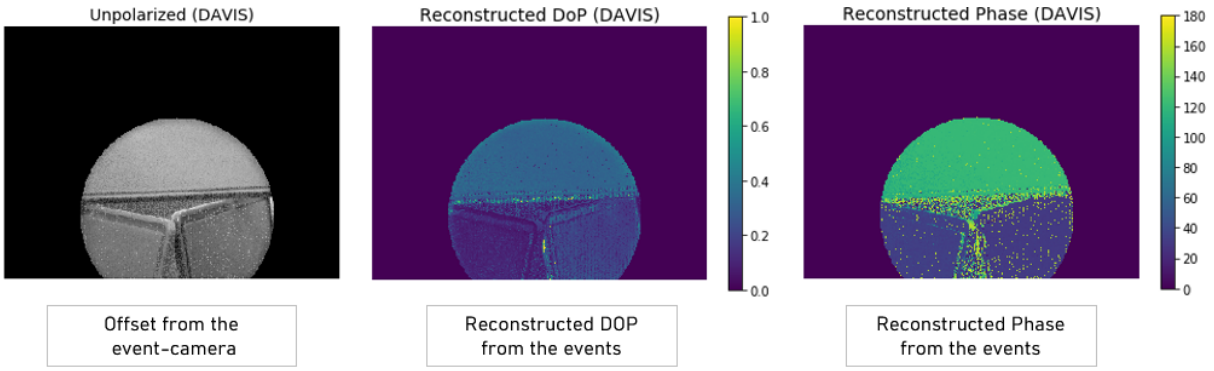


Figure 44: Encoded surface normal reconstructed from scene B.



(a)



(b)

Figure 45: Image representation of the DOP and phase the from the Basler ground-truth TRS samples (a), and from the reconstructed TRS samples (b) with the method proposed in subsection 6.2.2. Scene C.

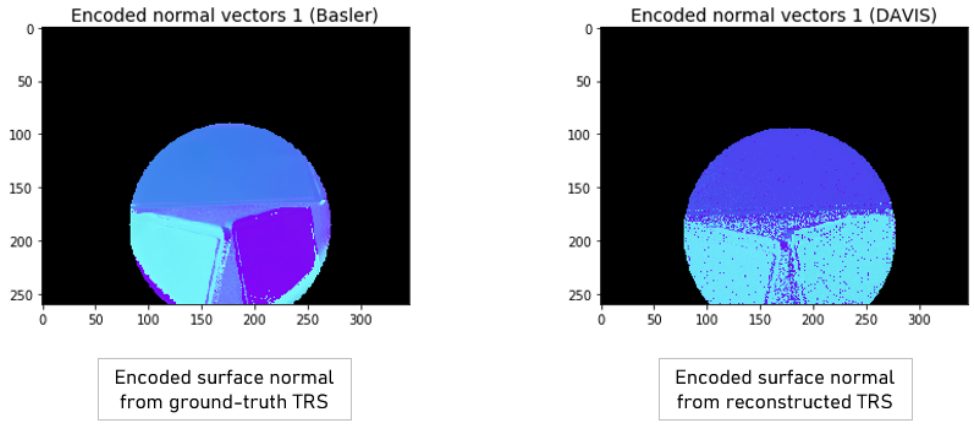


Figure 46: Encoded surface normal reconstructed from scene C.

From the the RGB-encoded surface normals of the scene of Figure 42, by taking as reference the encoded normal vectors computed from the polarization images captured with the Basler monochrome camera, the normal reconstructions from the DAVIS data does not perfectly match with the expected ones. By comparing the DOP and the TRS phase images of Figure 41, the reconstructed DOP does not match for the majority of the DAVIS pixels, while the reconstructed phase is slightly mismatching. Since the offset images are very similar when compared together, the higher DOP computed from the TRS reconstructions from the event-camera is related to an overestimation of the reconstructed amplitude. This may also be related to some non-idealities of the DAVIS camera [19]. Usually, once the event thresholds are set on the DAVIS sensor, the threshold value is not the same among the pixels but has some little variance, leading the DAVIS pixels not to all behave the same way. In fact, it has been noticed that there are some pixels that do not fire events that are somehow correlated to the sinusoid (for example only positive events are produced during the acquisition) and the network may not be able to perfectly reconstruct the sinusoid when we have these bad behaving pixels.

In the scene in figure 44, the reconstructed surface normal orientations from the DAVIS events are in general well matching with the expected orientations computed from the polarization images acquired with the Basler camera, except for some pixels capturing the object placed on the right side of the scene. In Figure 43, by comparing the gray-scale images captured with Basler and DAVIS cameras it is possible to see that the intensities of the object placed at the right side of the scene do not match for some pixels. Obviously, the pixels having a mismatching intensity have also different TRS offset when using the method presented in subsection 6.2.2 for the TRS offset reconstruction, leading again to a wrong DOP estimation from the reconstructed TRS. The mismatching intensities between the Basler monochrome camera and the the DAVIS monochrome acquisition may be due to the noise produced by the DVS pixels circuitry of the DAVIS camera when using the DAVIS APS pixels for acquiring gray-scale images [32], since by construction of the DAVIS sensor it is not possible to acquire gray-scale images and fully disable the DVS circuitry of the sensor. This noisy effect on the DAVIS grayscale images is also one of the reasons why this project relies on the Basler monochrome camera for acquiring the ground-truth data.

The last scene proposed in Figure 45 instead emphasizes a problem related to the phase estimation from the reconstructed TRS. For the object placed on the bottom-right side of the scene, the phase estimated from the reconstructed TRS is totally wrong, making the surface normal orientations of the objects placed at the bottom side of the

scene to point exactly in the same direction, as possible to see from the reconstructed normals in Figure 46. The wrong estimation of the phase is due to the fact that just a slightly forward shift of the reconstructed TRS samples with respect to the ground-truth samples leads to an abrupt phase shift of the sinusoid from $180°$ to $0°$, making then the reconstructed surface normal orientation to be horizontally flipped as illustrated in Figure 47.

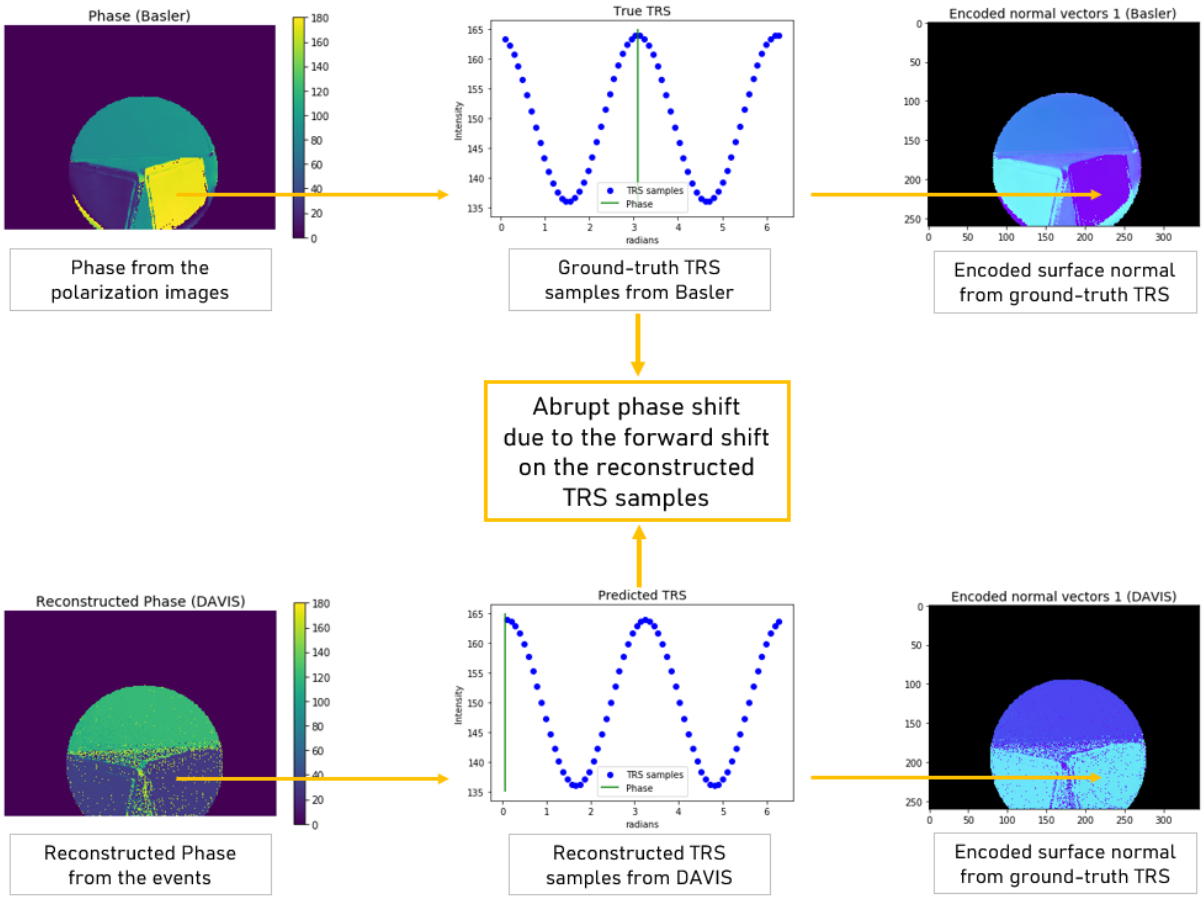


Figure 47: Figure showing the abrupt phase shift computed from the reconstructed TRS samples being slightly shifted forward.

7 Conclusions

With this work, a preliminary investigation on the possibility of performing surface normal reconstruction of scenes by combining a linear polarizing filter and an event-sensing camera has been carried out. A deep learning approach presented in subsection 6.2 has been developed in order to pixel-wise transform the events collected from an event-sensing camera to a representation resembling the signal acquired from a traditional monochrome camera, then the surface normals orientations are reconstructed according to the method proposed by [4]. Due to the lack of an already existing dataset which could have been useful for this project, the laboratory setup presented in subsection 5.1 has been settled for acquiring data from polarization cues in a real scenario from both an event-sensing camera and a monochrome camera, in order to produce a dataset for training in a fully supervised way different neural networks. Among some neural network tested, the 1D-Unet combined with the method presented in subsection 6.2.2 has produced the best results for the pixel-wise transformation of the the polarization data collected with the event-sensing camera. However, this method is still not perfect and needs to be furthermore revised due to some imperfections as presented in section 6.3. In fact, just a slight mismatch between the output of the network and the ground-truth data can lead to a totally wrong reconstruction of the surface normals orientation. Further developments could consists in producing a larger dataset by also collecting data with a faster rotation of the linear polarizing filter, which can lead to better results in the data acquisition with seethe event sensor, as discussed in subsection 4.4.
Nevertheless, it is important to mention that during this first investigation the event-sensing camera seems to have some issues when used for collecting data related to polarization cues for surface normal reconstruction, despite the advantages of event-sensing cameras over the traditional monochrome cameras. In fact, for this task, the event camera seems to be a very noisy sensor for collecting events related to polarization: when setting up the parameters of the event camera for collecting data related to polarization information of the light, a lot of noise is produced by the sensor due to its non-idealities [19].

List of Figures

Bibliography

[1] Lawrence B Wolff and Terrance E Boult. Constraining object features using a polarization reflectance model. *Phys. Based Vis. Princ. Pract. Radiom*, 1:167, 1993.

[2] Lawrence B Wolff, Todd A Mancini, Philippe Pouliquen, and Andreas G Andreou. Liquid crystal polarization camera. *IEEE transactions on Robotics and Automation*, 13(2):195–203, 1997.

[3] Yilbert Giménez, Pierre-Jean Lapray, Alban Foulonneau, and Laurent Bigué. Calibration algorithms for polarization filter array camera: survey and evaluation. *Journal of Electronic Imaging*, 29(4):041011, 2020.

[4] Gary A Atkinson and Edwin R Hancock. Recovery of surface orientation from diffuse polarization. *IEEE transactions on image processing*, 15(6):1653–1664, 2006.

[5] Gary A Atkinson and Edwin R Hancock. Surface reconstruction using polarization and photometric stereo. In *International conference on computer analysis of images and patterns*, pages 466–473. Springer, 2007.

[6] Trung Ngo Thanh, Hajime Nagahara, and Rin-ichiro Taniguchi. Shape and light directions from shading and polarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2310–2318, 2015.

[7] William AP Smith, Ravi Ramamoorthi, and Silvia Tozza. Linear depth estimation from an uncalibrated, monocular polarisation image. In *European Conference on Computer Vision*, pages 109–125. Springer, 2016.

[8] Fotios Logothetis, Roberto Mecca, Fiorella Sgallari, and Roberto Cipolla. A differential approach to shape from polarisation: A level-set characterisation. *International Journal of Computer Vision*, 127(11):1680–1693, 2019.

[9] Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Linares-Barranco, and Tobi Delbruck. Retinomorphic event-based vision sensors: bioinspired cameras with spiking output. *Proceedings of the IEEE*, 102(10):1470–1484, 2014.

[10] Shih-Chii Liu and Tobi Delbruck. Neuromorphic sensory systems. *Current opinion in neurobiology*, 20(3):288–295, 2010.

[11] Kunihiko Fukushima, Yoko Yamaguchi, Mitsuru Yasuda, and Shigemi Nagata. An electronic model of the retina. *Proceedings of the IEEE*, 58(12):1950–1951, 1970.

[12] Carver A Mead and Misha A Mahowald. A silicon model of early visual processing. *Neural networks*, 1(1):91–97, 1988.

[13] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 $120db$ $15\mu s$ latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008.

[14] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *IEEE Journal of Solid-State Circuits*, 46(1):259–275, 2010.

[15] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240$\times$ 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.

[16] Raphael Berner, Christian Brandli, Minhao Yang, S-C Liu, and Tobi Delbruck. A 240x180 120db 10mw 12us-latency sparse output vision sensor for mobile applications. In *Proceedings of the International Image Sensors Workshop*, number CONF, pages 41–44, 2013.

[17] DAVIS-346 RED specifications: `https://inivation.com/wp-content/uploads/2020/09/DAVIS346.pdf`.

[18] jAER framework information: `http://jaerproject.org`.

[19] Yuji Nozaki and Tobi Delbruck. Temperature and parasitic photocurrent effects in dynamic vision sensors. *IEEE Transactions on Electron Devices*, 64(8):3239–3245, 2017.

[20] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *CoRR*, abs/1904.08405, 2019.

[21] Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. Comparison of spike encoding schemes in asynchronous vision sensors: Modeling and design. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2632–2635. IEEE, 2014.

[22] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, pages 308–324. Springer, 2018.

[23] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126(12):1381–1393, 2018.

[24] Souptik Barua, Yoshitaka Miyatani, and Ashok Veeraraghavan. Direct face detection and video reconstruction from event cameras. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

[25] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019.

[26] Jonghyun Choi, Kuk-Jin Yoon, et al. Learning to super resolve intensity images from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2768–2776, 2020.

[27] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 156–163, 2020.

[28] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. Spade-e2vid: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Transactions on Image Processing*, 30:2488–2500, 2021.

[29] RAL-Plastics website: `https://ral-shop.com/product-category/ral-plastics/`.

[30] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[32] Christian P Brändli. *Event-based machine vision*. PhD thesis, ETH Zurich, 2015.